Outlier Detection using Inductive Logic Programming

Fabrizio Angiulli and Fabio Fassetti
DEIS, University of Calabria
87036 Rende (CS), Italy
{f.angiulli, f.fassetti}@deis.unical.it

Abstract—We present a novel definition of outlier in the context of inductive logic programming. Given a set of positive and negative examples, the definition aims at singling out the examples showing anomalous behavior. We note that the task here pursued is different from noise removal, and, in fact, the anomalous observations we discover are different in nature from noisy ones. We discuss peculiarities of the novel approach, present an algorithm for detecting outliers, discuss some examples of knowledge mined, and compare it with alternative approaches.

I. INTRODUCTION

Traditional approaches to outlier detection model the normal behavior of individuals by performing some statistical kind of computation on the given data set and, then, single out those individuals whose behavior or characteristics significantly deviate from normal ones. However, a very interesting direction of research concerns the capability of exploiting domain knowledge in order to guide search for anomalous observations. Indeed, while looking over a set of observations to discover outliers, it often happens that there is some qualitative description of the domain of interest encoding what an expected normal behavior should be. This description might be, for instance, derived by an expert and might be formalized by means of a suitable language for knowledge representation. With this aim, in [1], [2] a concept of outlier in the context of default reasoning ad logic programming is presented.

Inductive Logic Programming (ILP) is an important field at the intersection of machine learning and logic programming which aims at inducing relation descriptions of data in the form of logic programs [3]. In the context above delineated, we present a definition of outlier in the field of ILP. Given a set of positive and negative examples from a concept, the definition aims at singling out the examples showing exceptional behavior. The method is unsupervised, since there are no examples of normal/abnormal behavior, even if it has connections with supervised learning, since it is based on induction from examples which are instances of a concept. In particular, the definition distinguishes among three kinds of abnormalities, that are irregular, anomalous, and outlier observations. This allows us to provide a finer characterization of the anomaly at hand and to single out more subtle forms of anomalies. Moreover, we are also able to provide explanations for the abnormality of the observation, in the form of a pair of logic programs, which make more intelligible the motivation underlying its exceptionality.

ILP learning systems usually have a single mechanism, called noise-handling mechanism [4], for dealing with noisy, incomplete and inexact data, which prevents the induced hypothesis from overfitting the data set. While the presence of noise in the examples has some relationship with the approach here pursued, the other kinds of imperfect data are orthogonal to the present task, in that they are problems of the learning task in its entirety.

There are major differences also between the approach presented here and those introduced in [1], [2]. First, our approach is based of induction, while the others are based on deduction. Indeed, the disagreement of the abnormal observations with the theory at hand is perceived here by means of a measure of the difference of the generalization of the hypotheses induced in presence/absence of the observations (the compliance relationship), while in [1], [2] it is perceived by means of the satisfaction of certain conditions involving the entailment operator. Moreover, importantly, the definitions proposed in [1], [2] strongly rely on the non-monotonicity of the employed formalisms, which use either default rules or negation by default (other than the classical negation). As a matter of fact, if the logic program under consideration is positive, that is under the formal framework considered here, according to the definition provided in [2] there are no outliers in a logic program (cf. Theorem 3.2 of [2]). This makes the two approaches for defining outliers incomparable even from a practical point of view.

The rest of the paper is organized as follows. Section II presents some preliminary definitions. Section III introduces the notion of abnormal example. Section IV provides the definition of ILP-based outlier mining problem. Section V reports experimental results. Finally, Section VI draws conclusions.

II. PRELIMINARY DEFINITIONS

The reader is referred to [5] for basics on Logic Programming and to [3] for Inductive Logic Programming. Next we provide some definitions useful to define the concept of outlier in ILP.

The set of ground atoms covered by a clause \( c \) is denoted as \( \text{covers}(c) \). If \( C \) is a set of clauses, \( \text{covers}(C) \) is the union
of the set of ground atoms covered by the clauses in $C$. Let $E$ be a set of ground atoms. In the following, $\text{covers}_E(C)$ denotes the set $\text{covers}(C) \cap E$. By definition, we assume that $\text{covers}_E(\emptyset) = E$.

Given a logic program $P$ and a set of ground atoms $E$, the restriction $P(E)$ of $P$ to $E$ is the logic program $P(E) = \{ e \in P : e \in \text{covers}_E(C) \} \neq \emptyset$.

Let $U$ be an universal set of observations, also called objects or instances. A (direct) concept $C$ is a subset of $U$. The dual concept $\overline{C}$ of $C$ is the concept $U \setminus C$.

A set of examples $E$ is a set of ground atoms that can be partitioned in two subsets, that are $E^+$, the set of positive examples, and $E^-$, the set of negative examples.

The problem that ILP is interested in solving can be stated as follows: Given a set of examples $E$, find a hypothesis $\mathcal{H}_E$ such that $\mathcal{H}_E \cup \mathcal{B}$ entails the examples in $E$; namely:

1) for each $e \in E^+$, $e \in \text{covers}(\mathcal{H}_E \cup \mathcal{B})$ (completeness),
2) for each $e \in E^-$, $e \notin \text{covers}(\mathcal{H}_E \cup \mathcal{B})$ (consistency).

Next we provide the concepts of coverage, gain, and compliance.

Definition 1 (Coverage): Let $C$ be a set of clauses and let $E$ be a set of examples. Then the coverage $\text{cov}_E(C)$ of $C$ in $E$ is the following function:

$$
\text{cov}_E(C) = \frac{1}{|E|} \left( \prod_{c \in E} |\text{covers}_E(c)| \right)^{\frac{1}{|E|}}.
$$

Intuitively, the coverage of a set of clauses in a set of examples measures how many examples of the set are covered in average by the clauses. In particular, the definition of coverage here provided is based on the geometric mean in order to penalize the presence of rules covering few examples. We will employ the coverage as a measure of the generalization of a set of clauses.

Definition 2 (Gain): Given two sets of clauses $C_1$ and $C_2$ and a set of examples $E$, the gain $\text{gain}_E(C_1, C_2)$ of $C_1$ over $C_2$ in $E$ is defined as

$$
\text{gain}_E(C_1, C_2) = \text{cov}_E(C_1) - \text{cov}_E(C_2).
$$

A positive gain means that the clauses in $C_1$ averagely cover a larger number of examples in $E$ than the clauses in $C_2$. Intuitively, this means that the clauses in $C_1$ can be considered more general than those in $C_2$.

Given a set of examples $E$ and a nonempty subset $O$ of $E$ we say that $O$ is pure if either $O \subseteq E^+$ or $O \subseteq E^-$ hold.

Definition 3 (Compliance): Given a background knowledge $\mathcal{B}$, a set of examples $E$, and a pure subset $O$ of $E$, we say that $O$ $\alpha$-complies (or, simply, complies) with $E \cup \mathcal{B}$, written $E \cup \mathcal{B} \sim O$, if

$$
\text{gain}_{E^+ \setminus O}(\mathcal{H}_E \cup \mathcal{B}, \mathcal{H}_B) < \alpha,
$$

where $\alpha$ is an user-provided parameter in $[0, 1]$. Otherwise, $O$ does not comply with $E \cup \mathcal{B}$, written $E \cup \mathcal{B} \not\sim O$.

Intuitively, if a subset of examples does not comply with a background theory and the whole set of examples, then this means that the hypothesis induced in absence of this subset is significatively more general than the hypothesis induced when the examples are seen.

Given a set of examples $E$, the dual set $\overline{E}$ of $E$ is the set of examples $E$ such that $\overline{E}^+ = E^-$ and $\overline{E}^- = E^+$. Note that by using $\overline{E}$ as set of examples, the dual concept $\overline{C}$ of $C$ is learned, where $C$ is the concept of which the examples $E$ are instances. Let $p$ denote the target predicate symbol. When the dual concept is learned we will employ in the induced hypothesis the predicate symbol $\text{not} \_ p$ instead of $p$.

III. ABNORMAL EXAMPLES

Given a pure subset of examples $O$ of $E$, we argue that the compliance of these examples with $E \cup \mathcal{B}$ and $\overline{E} \cup \mathcal{B}$ can be exploited in order to understand if the set $O$ contains abnormal observations. In order to illustrate the concepts that will be defined in this section, we make use of an example.

Example 1. Let

$$
E^+ = \{ t_p(a), t_p(b), t_p(c), t_p(d), t_p(e), t_p(f), t_p(g) \},
$$

$$
E^- = \{ t_p(f), t_p(g), t_p(h), t_p(i), t_p(j), t_p(k), t_p(m) \},
$$

be a given set of examples, and

$$
\mathcal{B} = \{ p(o), p(x), p(y), p(z), q(a), q(b), q(c), q(l), q(m),
$$

$$
\{ f, r(g), r(n), s(e), s(f), s(g), s(h), s(n),
$$

$$
t(c), t(d), u(a), u(e), u(o), v(b), v(j), v(k),
$$

$$
w(m), w(l), w(x), w(y), w(z) \}
$$

be a given background knowledge. Figure 1 shows the examples in $E$ and the subsets covered by each predicate in the background theory.

Let the induced hypothesis $\mathcal{H}_B$ be (throughout the paper, between angle brackets we report the number of examples covered by each clause):

$$
\mathcal{H}_B = \begin{cases}
\{ c_1 \equiv t_p(X) \leftarrow q(X) \} & \text{(5)} \\
\{ c_2 \equiv t_p(X) \leftarrow w(X) \} & \text{(5)} \\
\{ c_3 \equiv t_p(X) \leftarrow u(X) \} & \text{(3)} \\
\{ c_4 \equiv t_p(X) \leftarrow t(X) \} & \text{(2)} 
\end{cases}
$$

Figure 1. Example theory.
Indeed, we call these examples difficult to discriminate them from a non-instance. Hence, cov $O$ be covered since we can imagine they are “far away” the dual description of the concept is not affected by the set

The induced dual hypothesis $H^c_{B}$ be:

\[
H^c_{B} = \begin{align*}
\tau_1 & \equiv \text{not } t_{p}(X) \leftarrow \tau(X) \\
\tau_2 & \equiv \text{not } t_{p}(h) \\
\tau_3 & \equiv \text{not } t_{p}(j) \\
\tau_4 & \equiv \text{not } t_{p}(k)
\end{align*}
\]

For the sake of simplicity, assume that $O$ is a subset of $E^+$. As already said, if the set $O$ does not comply with $E \cup B$, then the description of the concept would be significantly more concise if each example in $O$ were not observed. Hence, intuitively, we can say that the examples in $O$ are likely to do not match regularities joining the remaining instances of the concept. In order to better understand the kind of irregularity represented by the examples in $O$, the compliance of $O$ with $E \cup B$ has to be investigated. In particular, if $O$ does not comply with $E \cup B$, but $\overline{O}$ complies with $E \cup B$, then the description of the concept would be significantly more concise if each example in $O$ were not observed, whereas the dual description of the concept is not affected by the set of examples $O$. Thus, in this case the examples are hard to be covered since we can imagine they are “far away” the majority of the positive examples, but anyway “far” form the negative ones. We identify these examples as irregular.

**Definition 4 (Irregular set):** Given a background knowledge $B$, a set of examples $E$, and a subset $O$ of $E^+$ ($E^-$, resp.), we say that $O$ is irregular in $E \cup B$ if $E \cup B \not\models O$ ($E \cup B \models O$, resp.) and $E \cup B \models \overline{O}$ ($E \cup B \not\models \overline{O}$, resp.).

**Example 1 (continued).** Assume $a$ is set to 0.04 and consider the set $O = \{t_p(d_1)\}$. We note the set $O$ is irregular. Indeed, $E \cup B \not\models \{t_p(d_1)\}$, since if $t_p(d_1)$ were not seen then the induced theory $H^c_{B,O}$ would not contain $c_3$ being, hence, more concise. In particular, $cov_{E \setminus O}(H^c_{B,O}) = \frac{\frac{5}{6} \times \frac{3}{10}}{\frac{5}{6}} = 0.27$, while $cov_{E \cup O}(H^c_{B,O}) = \frac{\frac{5}{6} \times \frac{3}{10}}{\frac{5}{6}} = 0.37$, and the gain is 0.10. Conversely, $E \cup B \models \{t_p(d_1)\}$ since $H^c_{B,O}$ is not affected by the absence of $\{t_p(d_1)\}$ (the gain is zero).

A similar line of reasoning can be employed if $\overline{O}$ does not comply with $E \cup B$. In particular, if $\overline{O}$ does not comply with $E \cup B$ and $O$ complies with $E \cup B$, then the examples in $O$ well fit the concept to be learned, but they also have some commonalities with the dual concept, so that it is very difficult to discriminate them from a non-instance. Hence, we call these examples anomalous.

**Definition 5 (Anomalous set):** Given a subset $O$ of $E^+$ ($E^-$, resp.), we say that $O$ is anomalous in $E \cup B$ if $E \cup B \not\models O$ ($E \cup B \models O$, resp.) and $E \cup B \not\models \overline{O}$ ($E \cup B \models \overline{O}$, resp.).

**Example 1 (continued).** Consider now the set $O = \{t_p(b)\}$. This set is anomalous. Indeed, $E \cup B \not\models \{t_p(b)\}$ since the hypothesis $H^c_{B,O}$ induced in absence of $O$ coincides with $H^c_{E}$. Conversely, $E \cup B \models \overline{O}$ ($E \cup B \not\models \overline{O}$) if $\{t_p(b)\}$ since if $t_p(b)$ were not seen then the induced dual theory $H^c_{B,O}$ would contain the clause $t_p(X) \leftarrow \tau(X)$ instead of the facts $\tau_3$ and $\tau_4$. In fact, $cov_{E \setminus \overline{O}}(H^c_{B,O}) = \frac{\frac{5}{6} \times \frac{3}{10}}{\frac{2}{6}}$, $cov_{E \cup \overline{O}}(H^c_{B,O}) = \frac{\frac{5}{6} \times \frac{3}{10}}{\frac{2}{6}}$, and the gain is 0.08.

Assume now that $O$ does not comply with $E \cup B$ and also $\overline{O}$ does not comply with $E \cup B$. In this case also the description of the dual concept would be significantly more concise if each example in $\overline{O}$ were not observed. Intuitively, this means that the examples in $O$ present some commonalities with the non-instances of the concept to be learned. In this case, both the description of the concept $E$ and the description of the dual concept $\overline{E}$ would benefit if the examples in $O$ and in $\overline{O}$, respectively, were not observed. We can imagine that these examples are hard to be covered, since they lie either very close or even within the “shape” of the dual concept, and we identify these examples as outliers.

**Definition 6 (Outlier set):** Given a pure subset $O$ of $E$, we say that $O$ is outlier in $E \cup B$ if $E \cup B \not\models O$ and $E \cup B \models \overline{O}$.

**Example 1 (continued).** The set $O = \{t_p(e)\}$ is an outlier. Indeed, $E \cup O \models \{t_p(e)\}$, since if $t_p(e)$ were not seen then the induced theory $H^c_{B,O}$ would not contain $c_3$, being, hence, more concise. In particular, $cov_{E \setminus O}(H^c_{B,O}) = \frac{\frac{5}{6} \times \frac{3}{10}}{\frac{5}{6}}$, since $H^c_{B,O}$ would contain $c_3 \equiv t_p(X) \leftarrow p(X)$ instead of $c_2$ and $c_3$, $cov_{E \cup O}(H^c_{B,O}) = \frac{\frac{5}{6} \times \frac{3}{10}}{\frac{5}{6}}$ and the gain is 0.08. Moreover, $E \cup B \models \{t_p(e)\}$, since if $t_p(e)$ were not seen then the induced dual theory $H^c_{B,O}$ would contain the clause $t_\alpha \equiv \text{not } t_p(X) \leftarrow s(X)$ instead of $\tau_3$ and $\tau_4$. In particular, $cov_{E \setminus \overline{O}}(H^c_{B,O}) = \frac{\frac{5}{6} \times \frac{3}{10}}{\frac{2}{6}}$ and $cov_{E \cup \overline{O}}(H^c_{B,O}) = \frac{\frac{5}{6} \times \frac{3}{10}}{\frac{2}{6}}$, and the gain is 0.045.

A pure set of examples $O$ such that both $E \cup B \models O$ and $E \cup B \models \overline{O}$ hold is said to be normal, otherwise it is said to be abnormal. Abnormal set of examples can be partitioned into outlier, irregular and anomalous examples, according to what aforesaid. Table I summarizes the different kinds of abnormal example sets that have been defined.

### Table I

<table>
<thead>
<tr>
<th>$O \subset E$ pure</th>
<th>$E \cup B \models \overline{O}$</th>
<th>$E \cup B \not\models \overline{O}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>normal</td>
<td>irregular (positive)</td>
<td>outlier</td>
</tr>
<tr>
<td>anomalous (negative)</td>
<td>irregular (negative)</td>
<td></td>
</tr>
<tr>
<td>outlter</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

IV. **STATEMENT OF THE PROBLEM AND ALGORITHM**

In this section we define the outlier detection problem in the context of ILP. First of all, we introduce an alternative notion of compliance which is based on the previous one, but presents some advantages which will be discussed next. Intuitively, the novel definition of compliance focuses on the portion of the theory involving only the examples in $O$.

For the sake of simplicity, let $O$ be a set of positive examples and consider the theory $H^c_{B}$ induced in presence of $O$. The set of clauses in $H^c_{B}$ can be partitioned in two groups, that are the clauses in $H^c_{B}(O)$ that cover some of the
examples in $\mathcal{O}$, and the remaining ones, that are the clauses in $\widehat{\mathcal{H}}_B \setminus \widehat{\mathcal{H}}_E(\mathcal{O})$. We call starting theory the former theory, since it builds on the examples in $\mathcal{O}$.

As for the set of examples, it can be partitioned in three sets, that are the examples in $\mathcal{O}$, the examples $\mathcal{O}$ not in $\mathcal{O}$ and covered only by clauses in $\widehat{\mathcal{H}}_B \setminus \widehat{\mathcal{H}}_E(\mathcal{O})$, and the remaining ones. Consider now the theory $\widehat{\mathcal{H}}_B \setminus \widehat{\mathcal{H}}_E(\mathcal{O})$ induced in absence of $\mathcal{O}$. In this case, the piece of theory which is affected by the absence of the examples in $\mathcal{O}$ is that composed of the clauses that cover some of the examples in $\mathcal{O}$ (recall that these examples are those not covered by the clauses in $\widehat{\mathcal{H}}_B \setminus \widehat{\mathcal{H}}_E(\mathcal{O})$). Hence, we call ending theory this theory.

We can now redefine the compliance by exploiting the gain of the ending theory over the starting theory in $\mathcal{E}^+ \setminus \mathcal{O}$.

**Definition 7 (Compliance):** Given a pure subset $\mathcal{O}$ of $\mathcal{E}$, let $\mathcal{E}^+ \setminus \mathcal{O}$ (the ending theory), and $\mathcal{E} \cup \mathcal{B}$ (the starting theory), we say that $\mathcal{O}$-complies (or, simply, complies) with $\mathcal{E} \cup \mathcal{B}$ if

$$\text{gain}_{\mathcal{E}} \left( \widehat{\mathcal{H}}_E(\mathcal{O}), \widehat{\mathcal{H}}_B(\mathcal{O}) \right) < \alpha,$$

where the theories $\widehat{\mathcal{H}}_B(\mathcal{O})$ (the starting theory) and $\widehat{\mathcal{H}}_E(\mathcal{O})$ (the ending theory), are defined as follows:

- if $\mathcal{O} \subseteq \mathcal{E}^+$, then $\widehat{\mathcal{H}}_B(\mathcal{O})$ is $\mathcal{H}_B(\mathcal{O})$ and $\widehat{\mathcal{H}}_E(\mathcal{O})$, where $\widehat{\mathcal{H}}_E(\mathcal{O})$ is the set of examples in $\mathcal{E}^+ \setminus \mathcal{O}$ covered only by clauses of $\mathcal{H}_B$ that cover some examples in $\mathcal{O}$;
- if $\mathcal{O} \subseteq \mathcal{E}^-$, then $\widehat{\mathcal{H}}_B(\mathcal{O})$ is $\mathcal{H}_B(\mathcal{O})$ and $\widehat{\mathcal{H}}_E(\mathcal{O})$ is $\mathcal{H}_B(\mathcal{O})$, where $\widehat{\mathcal{H}}_E(\mathcal{O})$ is the set of examples in $\mathcal{E}^+ \setminus \mathcal{O} = \mathcal{E}^+$ covered only by clauses of $\mathcal{H}_B(\mathcal{O})$ that cover some examples in $\mathcal{O}$.

**Example 1 (continued).** Consider the outlier set $\mathcal{O} = \{ t_p(e) \}$. Then the starting theory $\widehat{\mathcal{H}}_B(\mathcal{O})$ is $\mathcal{H}_B(\mathcal{O}) = \{ c_3 \}$ and, moreover, $\mathcal{O}$ is

$$\text{covers}_{\mathcal{E}} \left( \mathcal{H}_B(\mathcal{O}) \right) \setminus \text{covers}_{\mathcal{E}} \left( \mathcal{H}_B(\mathcal{O}) \right) \setminus \text{covers}_{\mathcal{E}} \left( \mathcal{E}^+ \setminus \mathcal{O} \right)$$

that is $\{ t_p(e) \}$. Let the theory $\mathcal{H}_B(\mathcal{O})$ induced in absence of $\mathcal{O}$ be

$$\mathcal{H}_B(\mathcal{O}) = \{ c_1 \equiv t_p(X) \iff q(X) \}.$$ 

Then the ending theory $\widehat{\mathcal{H}}_B(\mathcal{O})$ is $\mathcal{H}_B(\mathcal{O}) = \{ c_5 \}$, and the gain is $\frac{1}{2} - \frac{1}{1} = 0.20$

As for the dual concept, $\mathcal{O} = \{ t_p(e) \}$, let the theory $\mathcal{H}_B(\mathcal{O})$ induced in absence of $\mathcal{O}$ be

$$\mathcal{H}_B(\mathcal{O}) = \{ r_1 \iff \text{not}_p(f(X) \iff q(X)) \}.$$ 

Then the ending theory $\widehat{\mathcal{H}}_B(\mathcal{O})$ is $\mathcal{H}_B(\mathcal{O}) = \{ c_7 \}$, and the gain is $\frac{1}{2} - \frac{1}{1} = 0.20$

Now we discuss the advantages of the novel definition of compliance.

First, we note that the starting theory and the ending theory associated with the abnormal set of examples play the role of explanation for its abnormality. Indeed, since they represent the portion of knowledge which is affected by the presence/absence of the abnormal set, by comparing them the analyst can understand the motivation underlying the abnormality of the example set. With this aim, if $\mathcal{O}$ (resp.) does not comply with $\mathcal{E} \cup \mathcal{B}$, then we call direct (resp.) explanation the pair $(\widehat{\mathcal{H}}_B(\mathcal{O}), \widehat{\mathcal{H}}_E(\mathcal{O}))$, also said the direct (resp.) starting/ending theories associated with $\mathcal{O}$ (resp.) in $\mathcal{H}_B$ ($\mathcal{H}_E$, resp.).

Second, comparing only the pieces of theories related to the abnormal set is more meaningful than comparing the two full theories, and, moreover, this kind of comparison makes the definition less sensitive to “global changes”.

Now we are in the position of formally defining the ILP-based Outlier Detection problem.

**Definition 8:** Given a background knowledge $\mathcal{B}$, a set of examples $\mathcal{E}$, and a maximum size $k_{\text{max}}$, find the minimal abnormal subsets $\mathcal{O}$ of $\mathcal{E}$ of size not exceeding $k_{\text{max}}$, together with their associated explanations.

The pseudo-code of the ILP-based-outlier-detector algorithm for mining the ILP-based outliers is reported in Figure 2. The functions $Gain^+$ and $Gain^-$ are used to test compliance according to Definition 7.

**V. Experiments**

In this section we present some experiments conducted by using the proposed algorithm. We implemented the outlier detection algorithm in Yap Prolog on top of the P-Progol system\footnote{1} which is based on the PROGOL algorithm\footnote{2}.

**A. Zoo data set**

In this experiment we considered the Zoo data set from the UCI Machine Learning Repository\footnote{2}. This database contains instances associated with animals. Each instance consists of the animal name, the class which it belongs to (amphibian, bird, fish, invertebrate, insect, mammal, reptile), the number of legs and other boolean attributes. We built a background theory consisting of one unary predicate for each boolean attribute, and of the binary predicate legs. We used as target
predicate the binary predicate class. The set of positive examples consists of one hundred facts. The set of negative examples, consisting of six hundreds facts, has been obtained by associating each animal with the classes it does not belong to.

We executed the algorithm with $\alpha = 0.05$ and $k_{\text{max}} = 1$. Besides the facts in the induced hypothesis and the induced dual hypothesis, which are classified as irregular sets, the algorithm reported the following abnormal sets: $O_1 = \{\text{class(amphibian, newt)}\}$ as positive outlier, $O_2 = \{\text{class(insect, ladybird)}\}$ as positive anomalous, and $O_3 = \{\text{class(mammal, platypus)}\}$ as positive anomalous. Next we briefly comment on the knowledge discovered by the method.

**Newt:** The positive outlier set $O_1 = \{\text{class(amphibian, newt)}\}$ is a fact in the direct theory, while it has as dual explanation the dual starting theory $\overline{H}_{B}^{\mathcal{E}}(O_1)$:

\[
\text{not\_class}(\text{amphibian, newt}) \leftarrow \text{catsize}(X) \quad \text{(44)}
\]
\[
\text{not\_class}(\text{amphibian, X}) \leftarrow \text{legs}(2, X) \quad \text{(27)}
\]
\[
\text{not\_class}(\text{amphibian, tuatara}) \quad \text{(1)}
\]
\[
\text{not\_class}(\text{amphibian, scorpion}) \quad \text{(1)}
\]

and the dual ending theory $\overline{H}_{B}^{\mathcal{E}}(O_1)$:

\[
\text{not\_class}(\text{amphibian, X}) \leftarrow \text{tail}(X) \quad \text{(74)}
\]

with gain 0.11. From this explanation, it is clear that the newt is the only amphibian in the example set having the tail. As a matter of fact, it is the only amphibian of the Caudata order belonging to the set of examples, while all the other amphibians in the example set belong to the Anura order, which is characterized by the absence of tail.

**Ladybird:** The positive anomalous set $O_2 = \{\text{class(insect, ladybird)}\}$ has dual explanation the dual starting theory $\overline{H}_{B}^{\mathcal{E}}(O_2)$:

\[
\text{not\_class}(\text{insect, X}) \leftarrow \text{aquatic}(X) \quad \text{(35)}
\]
\[
\text{not\_class}(\text{insect, X}) \leftarrow \text{predator}(X) \quad \text{(54)}
\]

with gain 0.08. Indeed, among the insects present in the examples, that are the flea, gnat, honeybee, housefly, moth, termite, and wasp, the ladybird is the only predator.

**Platypus:** The positive anomalous set $O_3 = \{\text{class(mammal, platypus)}\}$ has dual explanation the dual starting theory $\overline{H}_{B}^{\mathcal{E}}(O_3)$:

\[
\text{not\_class}(\text{mammal, X}) \leftarrow \text{legs}(6, X) \quad \text{(10)}
\]
\[
\text{not\_class}(\text{mammal, X}) \leftarrow \text{feathers}(X) \quad \text{(20)}
\]
\[
\text{not\_class}(\text{mammal, X}) \leftarrow \text{eggs}(X), \text{toothed}(X) \quad \text{(19)}
\]
\[
\text{not\_class}(\text{mammal, X}) \leftarrow \text{eggs}(X), \text{legs}(0, X) \quad \text{(19)}
\]
\[
\text{not\_class}(\text{mammal, X}) \leftarrow \text{eggs}(X) \quad \text{(57)}
\]

with gain 0.09. The platypus is a well-known strange mammal, since the female lays eggs, although the newly hatched young are fed by the mother’s milk.

### B. Student Loan

Here we consider the Student Loan relational domain from the UCI Machine Learning Repository\(^3\). The target unary predicate no\_payment\_due(Person) is true for those people who are not required to repay a student loan. Auxiliary relations can be used to fully discriminate positive from negative instances. We executed the algorithm with $\alpha = 0.05$ and $k_{\text{max}} = 1$, with 78 positive examples and 34 negative examples, consisting of the students whose identifier number starts with 1. Next we briefly comment on the negative outlier set $\overline{O}_1 = \{\text{no\_payment\_due(student149)}\}$. It has as direct explanation the direct starting theory $\overline{H}_{B}^{\mathcal{E}}(O_1)$:

\[^{3}\text{http://archive.ics.uci.edu/ml.}\]
D. Comparison with DB-Outliers

and the direct ending theory \( \Pi_E^c(O_1) \):

\[
\text{no\_payment\_due}(X) \leftarrow \text{enrolled}(X, Y, 10) \quad (12)
\]

and the direct ending theory \( \Pi_E^c(O_1) \):

\[
\text{no\_payment\_due}(X) \leftarrow \text{male}(X), \text{enrolled}(X, Y, 10) \quad (6)
\]

\[
\text{no\_payment\_due}(\text{student165}) \quad (1)
\]

\[
\text{no\_payment\_due}(\text{student112}) \quad (1)
\]

\[
\text{no\_payment\_due}(\text{student196}) \quad (1),
\]

with gain 0.13, while it is a fact in the dual theory. The \text{student149} is strange, since it is the only enrolled in ten units which is required to repay a student loan.

C. Comparison with noise-handling mechanisms

In order to compare our approach with noise-handling mechanisms implemented in ILP learning systems, we ran P-Progol by setting the parameter \text{noise} (representing an upper bound on the number of negative examples allowed to be covered by an acceptable clause) to 5 and the parameter \text{min-}\text{pos} (representing a lower bound on the number of positive examples to be covered by an acceptable clause) to 2. According to the induced hypothesis the positive examples associated with \text{clam}, \text{crab}, \text{seawasp}, \text{slug}, \text{starfish}, \text{worm}, \text{pitviper}, \text{seasnake}, \text{slowworm}, \text{tortoise}, \text{tuatara} are classified as noise. Moreover, the unique negative example classified as noise is \text{class}(\text{amphibian}, \text{tuatara}).

The negative examples classified as noise, were reported as irregular singleton sets by our method since they do not comply with the set of examples, being facts in the complete and consistent direct hypothesis. As for the negative example \text{class}(\text{amphibian}, \text{tuatara}), it was also reported as an irregular singleton set by our method, since it does not comply with the dual set of examples, being a fact in the complete and consistent dual hypothesis. As for its compliance with the set of examples, the starting theory associated with the set \{\text{class}(\text{amphibian}, \text{tuatara})\} is \{\text{class}(\text{amphibian}, \text{toad}), \text{class}(\text{amphibian}, \text{frog}), \text{class}(\text{amphibian}, \text{newt})\}, while the corresponding ending theory is:

\[
\text{class}(\text{amphibian}, X) \leftarrow \text{eggs}(X), \text{toothed}(X), \text{legs}(4, X) \quad (3)
\]

with associated gain 0.02. Hence, for \( \alpha \leq 0.02 \) it would be recognized as an outlier set by our method. We note that our method found the outlier set \{\text{class}(\text{amphibian}, \text{newt})\}, and the anomalous sets \{\text{class}(\text{insect}, \text{ladybird})\} and \{\text{class}(\text{mammal}, \text{platypus})\}, for which there is no counterpart in the noise returned by P-Progol.

D. Comparison with DB-Outliers

We compared the method here presented with the distance-based outlier definition [7], by searching the distance-based outliers in the original Zoo data set. We used as outlier score the sum of the distances to the \( k \)-nearest neighbors [8], the Hamming function as distance measure, and we set both \( k \) (the number of nearest neighbors to consider) and \( n \) (the number of outliers to return) to 5 (that corresponds to the 5% of the positive examples). The following table reports the top-\( n \) distance-based outliers:

<table>
<thead>
<tr>
<th>Outlier (Score)</th>
<th>Nearest Neighbors</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. scorpion (25)</td>
<td>worm, slug, pitviper, clam, crab</td>
</tr>
<tr>
<td>2. seasnake (19)</td>
<td>pitviper, stingray, chub, herring, bass</td>
</tr>
<tr>
<td>3. tortoise (18)</td>
<td>tuatara, astrich, rhea, slowworm, wren</td>
</tr>
<tr>
<td>4. toad (17)</td>
<td>frog, newt, tuatara, worm, crab</td>
</tr>
<tr>
<td>5. pitviper (15)</td>
<td>slowworm, tuatara, seasnake, newt, kiwi</td>
</tr>
</tbody>
</table>

As for the comparison with our method, the singleton sets of positive examples associated with the distance-based outliers 2, 3, 4, and 5, are returned as irregular sets by our method, while, as for the outlier 1, some singleton sets of negative examples involving it are returned as irregular sets. The example \text{class}(\text{invertebrate}, \text{scorpion}) is not recognized as abnormal, since it shares with the octopus the property of having eight legs. It is clear, that the abnormal instances returned by the distance-based method are of different nature with respect to those returned by the approach here introduced. In particular, distance-based outliers are likely to correspond to irregular instances, since, intuitively, they are objects whose attribute-value pattern is shared less.

VI. Conclusions

In this work, we presented a novel approach to detect outliers which exploits domain knowledge. We introduced the novel definition of ILP-based outlier. We discussed some examples of knowledge mined, and compared it with alternative approaches. As future work, we are intended to investigate the computational complexity of the problem and to design more efficient mining algorithms for dealing with large data sets.

REFERENCES


