COPYRIGHT NOTICE

DOI: `http://dx.doi.org/10.1007/s10618-013-0310-5`.

Noname manuscript No.
(will be inserted by the editor)

Exploiting Domain Knowledge to Detect Outliers

Fabrizio Angiulli · Fabio Fassetti

Received: date / Accepted: date

Abstract We present a novel definition of outlier whose aim is to embed an available domain knowledge in the process of discovering outliers. Specifically, given a background knowledge, encoded by means of a set of first-order rules, and a set of positive and negative examples, our approach aims at singling out the examples showing abnormal behavior. The technique here proposed is *unsupervised*, since there are no examples of normal or abnormal behavior, even if it has connections with *supervised* learning, since it is based on induction from examples. We provide a notion of compliance of a set of facts with respect to a background knowledge and a set of examples, which is exploited to detect the examples that prevent to improve generalization of the induced hypothesis. By testing compliance with respect to both the direct and the dual concept, we are able to distinguish 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 subtle forms of anomalies. Moreover, we are also able to provide *explanations* for the abnormality of an observation which make intelligible the motivation underlying its exceptionality. We present both exact and approximate algorithms for mining abnormalities. The approximate algorithms improve execution time while guaranteeing good accuracy. Moreover, we discuss peculiarities of the novel approach, present examples of knowledge mined, analyze the scalability of the algorithms, and provide comparison with noise handling mechanisms and some alternative approaches.

A preliminary version of this article appears under the title "Outlier Detection using Inductive Logic Programming" in the Proceedings of the *IEEE International Conference on Data Mining* (ICDM), Miami, Florida, December 6-9, 2009 (Angiulli and Fassetti, 2009b).

F. Angiulli · F. Fassetti
DIMES, University of Calabria
Via P. Bucci, 41C
Tel.: +39-0984-494717
Fax: +39-0984-494713
E-mail: {f.angiulli, f.fassetti}@dimes.unical.it

Keywords Outlier Detection · Unsupervised methods · Knowledge Representation · Concept Learning

1 Introduction

Outlier detection is an active research field in data mining that has many applications in all those domains that can lead to illegal or abnormal behavior, such as fraud detection, network intrusion detection, medical diagnosis, marketing or customer segmentation (Hodge and Austin, 2004; Chandola et al, 2009).

Approaches to outlier detection can be classified in *supervised*, *semi-supervised*, and *unsupervised*. *Supervised* methods exploit the availability of a labeled data set, containing observations already labeled as normal and abnormal, in order to build a model of the normal class (Chawla et al, 2004). Since usually normal observations are the great majority, these data sets are unbalanced and specific classification techniques have to be designed to deal with the presence of rare classes. *Semi-supervised* methods assume that only normal examples are given. The goal is to find a description of the data, that is a rule partitioning the object space into an accepting region, containing the normal objects, and a rejecting region, containing all the other objects (Schölkopf et al, 1995). These methods are also called one-class classifiers or domain description techniques, and they are related to novelty detection since the domain description is used to identify objects significantly deviating form the training examples. *Unsupervised* methods search for outliers in an unlabelled data set by assigning to each object a score which reflects its degree of abnormality (Knorr and Ng, 1998; Breunig et al, 2000; Papadimitriou et al, 2003; Angiulli and Pizzuti, 2005; Kriegel et al, 2008; Angiulli and Fassetti, 2009a). Scores are usually computed by comparing each object with objects belonging to its neighborhood.

Traditional approaches 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 the 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 (Angiulli et al, 2007, 2008) a notion of outlier in the context of default reasoning and logic programming is presented, definition which exploits the deduction mechanism in order to single out outliers.

In the context above delineated, that is incorporating domain knowledge in the outlier mining task, we take the opposite perspective and present a definition of outlier which exploits induction in order to perceive the anomaly of an observation. Given a domain knowledge and a set of positive and negative

examples from a concept, the definition aims at singling out the examples showing exceptional behavior. The method is an *unsupervised* one, 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 an observation, in the form of a pair of logic programs, which make more intelligible the motivation underlying its exceptionality.

As a tool of concept learning based on logic programs (that is to say, set of first-order rules), we consider the field of Inductive Logic Programming (ILP), which aims at inducing descriptions of data in the form of logic programs (Lavrač and Džeroski, 1994). During recent years, ILP has shown its application potential in many fields, such as knowledge discovery in databases, relational data mining, bioinformatics, and others.

We present three algorithms for detecting Concept-Based outliers. The first algorithm, named CBOut, returns all and only the abnormal sets but may have large time requirements, while the other two algorithms, named hCBOut and aCBOut, greatly reduce execution time while guaranteeing good accuracy.

We discuss differences with noise-handling mechanisms, pointing out that the task here pursued is different from noise removal, since the anomalous observations we discover are different in nature from noisy ones. We also discuss differences with methods presented in (Angiulli et al, 2007, 2008), highlighting that the latter ones do not make sense in the context of positive logic programs, which is the framework here considered. Moreover, we discuss examples of knowledge mined and compare our approach with alternative ones.

The rest of the paper is organized as follows. Section 2 presents basic notions of Logic Programming and Concept Learning. Section 3 provides the formal definition of Concept-Based outlier. Section 4 describes the CBOut algorithm for mining outliers. Section 5 presents approximate variants of the CBOut algorithm for dealing with complex background theories and large set of examples. Section 6 discusses relationships and differences with related work. Section 7 reports results obtained by experimenting the approach here introduced. Finally, Section 8 draws conclusions of the work.

2 Preliminaries

In this section we recall some preliminary notions about Logic Programming, Concept Learning and Inductive Logic Programs.

2.1 Logic Programming

Next, we recall some basic concepts about *Logic Programming*.

A *term* is a constant or a variable. An *atom* is an expression of the form $p(t_1, ..., t_k)$ where p is a (k-ary) predicate symbol and $t_1, ..., t_k$ are terms.

A *Horn clause*, or simply *clause*, is an expression of the form $T \leftarrow Q$ where T is an atom and Q is a, possibly empty, conjunction of atoms, written $Q \equiv A_1, \ldots, A_n$, with A_i an atom ($n \geq 0$ and $1 \leq i \leq n$). If Q is empty ($n = 0$), then the clause is said a *fact*, also denoted simply as T.

A *Logic Program* P is a finite set of clauses.

The *Herbrand Universe* U_P of a program P is the set of all constants appearing in P, and its *Herbrand Base* B_P is the set of all ground atoms that can be obtained by combining the predicate symbols appearing in P and the constants from U_P. A *ground* term (resp. an atom, a clause or a program) is a term (resp. an atom, a clause or a program) where no variables occur. The set of all ground instances of the rules in P is denoted by $ground(P)$.

An *interpretation* of P is any subset of B_P. The truth value of an atom a w.r.t. an interpretation I, denoted $value_I(a)$, is $true$ if $a \in I$ and $false$ otherwise. The truth value of a conjunction of ground atoms $F \equiv a_1, \ldots, a_n$, denoted as $value_I(F)$, is $true$ if $value_I(a_i) = true$ for each $1 \leq i \leq n$, $false$ otherwise. If F is empty then $value_I(F) = true$.

A ground clause $c : T \leftarrow Q$ is *satisfied* by the interpretation I if (i) $value_I(Q) = false$ or (ii) $value_I(Q) = true$ and $value_I(T) = true$. Thus, if Q is empty then c is satisfied by I if $value_I(T) = true$. An interpretation $\mathcal{M}$ for P is a model of P if $\mathcal{M}$ satisfies all clauses in $ground(P)$.

A logic program P *entails* a ground atom f (or, equivalently, f is entailed by P), written $P \models f$, if and only if f is $true$ in each model of P. Otherwise, f is not entailed by P, written $P \not\models f$.

It can be shown that $\mathcal{M}_0^P = \bigcap_i \mathcal{M}_i^P$, the intersection of all models $\mathcal{M}_i^P$ of the logic program P, is a model of P too, called the *minimal model* of P (since no proper subset of $\mathcal{M}_0^P$ is a model of P). This model has the important property that $P \models f$, with f a ground atom, if and only if $f \in \mathcal{M}_0^P$ (Lloyd, 1987).

A *substitution* $\theta = \{X_1/t_1, \ldots, X_k/t_k\}$ is a mapping from variables to terms. The application of a substitution θ to a conjunction of atoms F, denoted as $F\theta$, is obtained by replacing each occurrence of the variable X_i in F with the term t_i, for each $1 \leq i \leq n$.

Let P be a logic program, a clause $c : T \leftarrow Q$ of P *covers* a ground atom f if there exists a substitution θ, such that $T\theta = f$ and $Q\theta \subseteq \mathcal{M}_0^P$, where $\mathcal{M}_0^P$ is the minimal model of P. The set of ground atoms covered by c is denoted as $covers(c)$. The coverage notion can be extended to a set of clauses: if C is a set of clauses, $covers(C)$ 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, $covers_E(C)$ denotes the set $covers(C) \cap E$. By definition, we assume that $covers_E(\emptyset) = E$.

Given a logic program P and a set of ground atoms E, the *restriction* $P(E)$ of P to E is the logic program $P(E) = \{c \in P \mid covers_E(\{c\}) \neq \emptyset\}$ composed of the clauses c in P such that $covers_E(\{c\})$ is not empty.

2.2 Concept Learning

Next, basic definitions about Concept Learning and Inductive Logic Programming (ILP) can be provided.

Let U be an universal set of *observations*, also called *objects* or *instances*. A *concept* $\mathcal{C}$ is a subset of U. A concept can be described *extensionally* (by listing its instances) or *intensionally* (by giving a concise description of the concept in terms of rules). The *dual concept* $\overline{\mathcal{C}}$ of $\mathcal{C}$ is the concept $U \setminus \mathcal{C}$. Sometimes we will call $\mathcal{C}$ the *direct concept* in order to differentiate it from its dual one $\overline{\mathcal{C}}$.

The *Inductive Concept Learning* aims to learn an intensional description of a specific concept by induction from some given instances and non-instances of the concept at hand. The instances of the concept to be learned are called *positive examples*, whereas the non-instances are called *negative examples*. The induced description is also called *hypothesis*.

The inductive learning may often exploit not only the examples but also some prior knowledge in order to build a more concise description of the concept. This knowledge is called *background knowledge* or, equivalently, *background theory*.

Inductive Logic Programming (ILP) is a branch of the inductive concept learning where the objects, the hypothesis and the background knowledge are all expressed in terms of logic programs.

In particular, the concept to be learned is a predicate, referred to as *target predicate*, the objects are ground facts, the background knowledge and the hypothesis to be induced are logic programs.

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

For example, consider the concept "transport by land" encoded by means of the unary target predicate transport_by_land. The following is a set of positive and negative examples pertaining to this concept:

$$\mathcal{E}^+ = \begin{cases} transport\_by\_land(bike). \\ transport\_by\_land(motorbike). \\ transport\_by\_land(car). \\ transport\_by\_land(jeep). \\ transport\_by\_land(truck). \\ transport\_by\_land(bus). \\ transport\_by\_land(hovercraft). \end{cases}$$

$$\mathcal{E}^- = \begin{cases} transport\_by\_land(airplane). \\ transport\_by\_land(seaplane). \\ transport\_by\_land(airship). \\ transport\_by\_land(helicopter). \end{cases}$$

As already noticed, other than a set of examples, usually also a background knowledge $\mathcal{B}$ is available. A possible background knowledge associated with the

"transport by land" concept is reported next.

$$\mathcal{B} = \begin{cases} has\_propeller(hovercraft). \\ has\_propeller(airplane). \\ has\_propeller(seaplane). \\ has\_propeller(helicopter). \\ has\_propeller(airship). \\ \\ has\_steering\_wheel(car). \\ has\_steering\_wheel(truck). \\ has\_steering\_wheel(bus). \\ has\_steering\_wheel(jeep). \\ \\ travels\_on\_wheels(motorbike). \\ travels\_on\_wheels(bike). \\ \\ vertical\_take\_off(helicopter). \\ vertical\_take\_off(airship). \\ \\ has\_wings(airplane). \\ has\_wings(seaplane). \\ \\ travels\_on\_wheels(X) \leftarrow has\_steering\_wheel(X). \end{cases}$$

The problem that ILP is interested in solving can be stated as follows: *Given a set of examples $\mathcal{E}$ and a background knowledge $\mathcal{B}$, find a hypothesis $\mathcal{H}_{\mathcal{B}}^{\mathcal{E}}$, also said hypothesis on $\mathcal{E}$ w.r.t. $\mathcal{B}$, such that $\mathcal{H}_{\mathcal{B}}^{\mathcal{E}} \cup \mathcal{B}$ entails the examples in $\mathcal{E}$; namely:*

1. *for each $e \in \mathcal{E}^+$, $e \in covers(\mathcal{H}_{\mathcal{B}}^{\mathcal{E}} \cup \mathcal{B})$ or, equivalently, $\mathcal{H}_{\mathcal{B}}^{\mathcal{E}} \cup \mathcal{B} \models e$ (completeness), and*
2. *for each $e \in \mathcal{E}^-$, $e \notin covers(\mathcal{H}_{\mathcal{B}}^{\mathcal{E}} \cup \mathcal{B})$, or, equivalently, $\mathcal{H}_{\mathcal{B}}^{\mathcal{E}} \cup \mathcal{B} \not\models e$ (consistency).*

It has been shown that the ILP problem is undecidable in the general case (Plotkin, 1971). Thus, in the literature, different heuristic systems have been introduced for determining a sub-optimal solution to the ILP problem, among the others, GOLEM (Muggleton and Feng, 1990), FOIL (Quinlan and Cameron-Jones, 1993), and PROGOL (Muggleton, 1995).

Consider again the concept "transport by land". Starting from the set of examples and the background theory above reported, the following hypothesis can be suitably induced:

$$\mathcal{H}_{\mathcal{B}}^{\mathcal{E}} = \begin{cases} transport\_by\_land(X) \leftarrow travels\_on\_wheels(X). \\ transport\_by\_land(hovercraft). \end{cases}$$

3 Detecting Outliers through Concept Learning

In this section we formally define the Concept-Based outlier detection problem.

3.1 Preliminary definitions

In order to define the outliers, we introduce some preliminary notions.

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

$$cov_{\mathcal{E}}(C) = \frac{1}{|\mathcal{E}|}\left(\prod_{c \in C} |covers_{\mathcal{E}}(c)|\right)^{\frac{1}{|C|}}. \quad (1)$$

By definition, we assume that $cov_{\mathcal{E}}(\emptyset) = 1$.

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 $\mathcal{E}$, the gain $gain_{\mathcal{E}}(C_1, C_2)$ of C_1 over C_2 in $\mathcal{E}$ is defined as

$$gain_{\mathcal{E}}(C_1, C_2) = cov_{\mathcal{E}}(C_1) - cov_{\mathcal{E}}(C_2).$$

A positive gain means that the clauses in C_1 averagely cover a larger number of examples in $\mathcal{E}$ 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 $\mathcal{E}$ and a nonempty subset $\mathcal{O}$ of $\mathcal{E}$ we say that $\mathcal{O}$ is *pure* if either $\mathcal{O} \subseteq \mathcal{E}^+$ or $\mathcal{O} \subseteq \mathcal{E}^-$ hold.

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

$$gain_{\mathcal{E}^+ \setminus \mathcal{O}}\left(\mathcal{H}_{\mathcal{B}}^{\mathcal{E} \setminus \mathcal{O}}, \mathcal{H}_{\mathcal{B}}^{\mathcal{E}}\right) < \alpha,$$

where α is an user-provided ral number in $[0, 1]$. Otherwise, we say that $\mathcal{O}$ does not comply with $\mathcal{E} \cup \mathcal{B}$, written $\mathcal{E} \cup \mathcal{B} \not\rightsquigarrow \mathcal{O}$.

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 significantly more general than the hypothesis induced when the examples are seen.

Now, we introduce the notion of *dual set of examples*. Given a set of examples $\mathcal{E}$, the dual set $\overline{\mathcal{E}}$ of $\mathcal{E}$ is the set of examples $\overline{\mathcal{E}}$ such that $\overline{\mathcal{E}}^+ = \mathcal{E}^-$ and $\overline{\mathcal{E}}^- = \mathcal{E}^+$. Note that by using $\overline{\mathcal{E}}$ as set of examples, the dual concept $\overline{\mathcal{C}}$ of $\mathcal{C}$ is learned, where $\mathcal{C}$ is the concept of which the examples in $\mathcal{E}$ are instances.

Let p denote the target predicate symbol. When the dual concept is learned, for the sake of clarity, we will employ in the induced hypothesis the predicate symbol *not_p* instead of p, in order to emphasize the fact that the dual concept has been learned.

3.2 Defining abnormal examples

Before providing the definition of abnormal observation, the intuition underlying the approach that will be pursued here is informally illustrated by means of an example.

Consider the concept "transport by land" described in Section 2.2 and, moreover, consider the set $\mathcal{O} = \{transport\_by\_land(hovercraft)\}$ consisting on the positive example $transport\_by\_land(hovercraft)$. As already shown before, the induced hypothesis associated with this concept is

$$\mathcal{H}_{\mathcal{B}}^{\mathcal{E}} = \begin{cases} transport\_by\_land(X) \leftarrow travels\_on\_wheels(X). \\ transport\_by\_land(hovercraft). \end{cases}$$

Noticeably, the example $transport\_by\_land(hovercraft)$ appears as a fact in $\mathcal{H}_{\mathcal{B}}^{\mathcal{E}}$, while the remaining part of the theory $\mathcal{H}_{\mathcal{B}}^{\mathcal{E}}$ consists of a single rule covering any other positive example.

Intuitively, this kind of knowledge suggests that, unlike the rest of the positive examples, the example $transport\_by\_land(hovercraft)$ is hard to be covered and, hence, does not comply very well with the normal behavior of the concept. As a matter of fact, among the examples appearing in the set of positive examples, the hovercraft is the only vehicle moving on land which is not equipped with wheels. Indeed, differently from any other vehicle traveling over land, the hovercraft is the only that exploits an air-cushion in order to move on surfaces. This kind of abnormal behavior is witnessed by the hypothesis induced in absence of the positive examples in $\mathcal{O}$, which is more compact than the original one, consisting of the following single rule:

$$\mathcal{H}_{\mathcal{B}}^{\mathcal{E}\setminus\mathcal{O}} = \Big\{ transport\_by\_land(X) \leftarrow travels\_on\_wheels(X).$$

A set of observations $\mathcal{O}$ like the one above commented on will be referred to as *irregular* in the following.

In order to properly characterize the exceptionality of the set $\mathcal{O}$, the dual concept "not transport by land" has to be taken into account. Specifically, the hypothesis $\mathcal{H}_{\mathcal{B}}^{\overline{\mathcal{E}}}$ induced on the target predicate $not\_transport\_by\_land$ by considering the dual set of examples $\overline{\mathcal{E}}$ is reported next:

$$\mathcal{H}_{\mathcal{B}}^{\overline{\mathcal{E}}} = \begin{cases} not\_transport\_by\_land(X) \leftarrow has\_wings(X). \\ not\_transport\_by\_land(X) \leftarrow vertical\_take\_off(X). \end{cases}$$

Indeed, the negative examples are vehicles traveling by air which can be partitioned in two sets: some of them have wings, while some others are aircrafts that take off and land vertically. Interestingly, also the dual hypothesis induced in absence of the examples in $\mathcal{O}$ is more compact than the original one, since it consists of the following single rule:

$$\mathcal{H}_{\mathcal{B}}^{\overline{\mathcal{E}}\setminus\overline{\mathcal{O}}} = \big\{ not\_transport\_by\_land(X) \leftarrow has\_propeller(X).$$

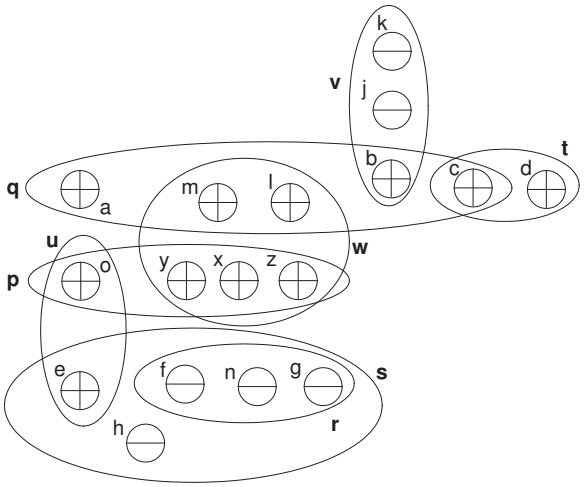


Fig. 1: Example theory.

As a matter of fact, while the hovercraft is the only positive example having a propeller, all the negative examples have a propeller. Hence, the presence of the example hovercraft prevents the above rule to be induced in the original dual theory. Intuitively, this kind of situation witnesses that $\mathcal{O}$ does not comply very well with the normal behavior of the dual concept, so that there are some difficulties to recognize the hovercraft as a non-instance of the concept "transport by land". In the following, we will call *anomalous* such an observation.

Having provided the intuition underlying the approach here pursued to perceive outliers, the rest of the section aims to define a way to measure the compliance of a set of examples to a hypothesis and to exploit it in order to properly formalize the notion of irregular and abnormal set of examples.

3.3 Abnormal examples

Given a pure subset of examples $\mathcal{O}$ of $\mathcal{E}$, we argue that the compliance of these examples with $\mathcal{E} \cup \mathcal{B}$ and $\overline{\mathcal{E}} \cup \mathcal{B}$ can be exploited in order to understand if the set $\mathcal{O}$ 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 $\mathcal{E} = \mathcal{E}^+ \cup \mathcal{E}^-$ be a given set of examples, where

$$\mathcal{E}^+ = \{t_p(a), t_p(b), t_p(c), t_p(d), t_p(e), t_p(l), t_p(m), t_p(o), t_p(x), t_p(y), t_p(z)\},$$
$$\mathcal{E}^- = \{t_p(f), t_p(g), t_p(h), t_p(j), t_p(k), t_p(n)\},$$

and let

$$\begin{aligned}\mathcal{B} = \{&p(o),\ p(x),\ p(y),\ p(z),\ q(a),\ q(b),\ q(c),\ q(l),\ q(m),\\ &r(f),\ r(g),\ r(n),\ s(e),\ s(f),\ s(g),\ s(h),\ s(n),\\ &t(c),\ t(d),\ u(e),\ u(o),\ v(b),\ v(j),\ v(k),\\ &w(m),\ w(l),\ w(x),\ w(y),\ w(z)\}\end{aligned}$$

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

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

$$\mathcal{H}_{\mathcal{B}}^{\mathcal{E}} = \begin{cases} c_1 \equiv t_p(X) \leftarrow q(X) & \langle 5 \rangle \\ c_2 \equiv t_p(X) \leftarrow w(X) & \langle 5 \rangle \\ c_3 \equiv t_p(X) \leftarrow u(X) & \langle 2 \rangle \\ c_4 \equiv t_p(X) \leftarrow t(X) & \langle 2 \rangle \end{cases}$$

and the induced dual hypothesis $\mathcal{H}_{\mathcal{B}}^{\overline{\mathcal{E}}}$ be:

$$\mathcal{H}_{\mathcal{B}}^{\overline{\mathcal{E}}} = \begin{cases} \overline{c}_1 \equiv not\_t_p(X) \leftarrow r(X) & \langle 3 \rangle \\ \overline{c}_2 \equiv not\_t_p(h) & \langle 1 \rangle \\ \overline{c}_3 \equiv not\_t_p(j) & \langle 1 \rangle \\ \overline{c}_4 \equiv not\_t_p(k) & \langle 1 \rangle \end{cases}$$

For the sake of simplicity, assume that $\mathcal{O}$ is a subset of $\mathcal{E}^+$. As already said, if the set $\mathcal{O}$ does not comply with $\mathcal{E} \cup \mathcal{B}$, then the description of the concept would be significantly more concise if each example in $\mathcal{O}$ were not observed. Hence, intuitively, we can say that the examples in $\mathcal{O}$ 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 $\mathcal{O}$, the compliance of $\overline{\mathcal{O}}$ with $\overline{\mathcal{E}} \cup \mathcal{B}$ has to be investigated. In particular, if $\mathcal{O}$ does not comply with $\mathcal{E} \cup \mathcal{B}$, but $\overline{\mathcal{O}}$ complies with $\overline{\mathcal{E}} \cup \mathcal{B}$, then the description of the concept would be significantly more concise if each example in $\mathcal{O}$ were not observed, whereas the dual description of the concept is not affected by the set of examples $\mathcal{O}$. 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 and also "far" form the negative ones. We identify these examples as *irregular*.

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

Example 1 (continued). Assume α is set to 0.05 and consider the set $\mathcal{O} = \{t_p(d)\}$. We note that the set $\mathcal{O}$ is irregular. Indeed, $\mathcal{E} \cup \mathcal{O} \not\rightsquigarrow \{t_p(d)\}$, since if $\{t_p(d)\}$ were not seen then the induced theory $\mathcal{H}_{\mathcal{B}}^{\mathcal{E}\setminus\mathcal{O}}$ would not contain c_4, being more concise. In particular, $cov_{\mathcal{E}^+\setminus\mathcal{O}}(\mathcal{H}_{\mathcal{B}}^{\mathcal{E}}) = \frac{\sqrt[4]{5\cdot5\cdot2\cdot1}}{10} = 0.27$, while $cov_{\mathcal{E}^+\setminus\mathcal{O}}(\mathcal{H}_{\mathcal{B}}^{\mathcal{E}\setminus\mathcal{O}}) = \frac{\sqrt[3]{5\cdot5\cdot2}}{10} = 0.37$, and the gain is 0.10.

Conversely, consider the set $\overline{\mathcal{O}}$ which is a subset of the negative examples in $\overline{\mathcal{E}}$ for the dual concept. Notice that $\overline{\mathcal{E}} \cup \mathcal{B} \rightsquigarrow \overline{\mathcal{O}}$, since $\mathcal{H}_{\mathcal{B}}^{\overline{\mathcal{E}}\setminus\overline{\mathcal{O}}}$ is not affected by the absence of $\overline{\mathcal{O}}$ (the gain is zero).

A similar line of reasoning can be employed if $\overline{\mathcal{O}}$ does not comply with $\overline{\mathcal{E}} \cup \mathcal{B}$. In particular, if $\overline{\mathcal{O}}$ does not comply with $\overline{\mathcal{E}} \cup \mathcal{B}$ and $\mathcal{O}$ complies with $\mathcal{E} \cup \mathcal{B}$, then the examples in $\mathcal{O}$ 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. In this case the description of the dual concept would be significantly more concise if each example in $\overline{\mathcal{O}}$ were not observed. Hence, we call these examples *anomalous*.

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

Example 1 (continued). Consider now the set $\mathcal{O} = \{t_p(b)\}$. This set is anomalous. Indeed, $\mathcal{E} \cup \mathcal{B} \rightsquigarrow t_p(b)$ since the hypothesis $\mathcal{H}_{\mathcal{B}}^{\mathcal{E}\setminus\mathcal{O}}$ induced in absence of $\mathcal{O}$ coincides with $\mathcal{H}_{\mathcal{O}}^{\mathcal{E}}$.

Conversely, consider the set $\overline{\mathcal{O}}$ which is a subset of the negative examples in $\overline{\mathcal{E}}$ for the dual concept. We note that $\overline{\mathcal{E}} \cup \mathcal{B} \not\rightsquigarrow \overline{\mathcal{O}}$, since if $\overline{\mathcal{O}}$ were not seen then the induced dual hypothesis $\mathcal{H}_{\mathcal{B}}^{\overline{\mathcal{E}}\setminus\overline{\mathcal{O}}}$ would contain the clause $not\_t_p(X) \leftarrow v(X)$ instead of the facts $\overline{c}_3$ and $\overline{c}_4$. In fact, $cov_{\overline{\mathcal{E}}^+}(\mathcal{H}_{\mathcal{B}}^{\overline{\mathcal{E}}}) = \frac{\sqrt[4]{3\cdot1\cdot1\cdot1}}{6}$, $cov_{\overline{\mathcal{E}}^+}(\mathcal{H}_{\mathcal{B}}^{\overline{\mathcal{E}}\setminus\overline{\mathcal{O}}}) = \frac{\sqrt[3]{3\cdot2\cdot1}}{6}$, and the gain is 0.08.

Assume now that $\mathcal{O}$ does not comply with $\mathcal{E} \cup \mathcal{B}$ and also that $\overline{\mathcal{O}}$ does not comply with $\overline{\mathcal{E}} \cup \mathcal{B}$. In this case, both the description of the concept $\mathcal{C}$ and the description of the dual concept $\overline{\mathcal{C}}$ would benefit if the examples in $\mathcal{O}$ and in $\overline{\mathcal{O}}$, 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 background knowledge $\mathcal{B}$, a set of examples $\mathcal{E}$, and a pure subset $\mathcal{O}$ of $\mathcal{E}$, we say that $\mathcal{O}$ is *outlier* in $\mathcal{E} \cup \mathcal{B}$ if $\mathcal{E} \cup \mathcal{B} \not\rightsquigarrow \mathcal{O}$ and $\overline{\mathcal{E}} \cup \mathcal{B} \not\rightsquigarrow \overline{\mathcal{O}}$.

Example 1 (continued). The set $\mathcal{O} = \{t_p(e)\}$ is an outlier. Indeed, $\mathcal{E} \cup \mathcal{B} \not\rightsquigarrow \{t_p(e)\}$, since if $\mathcal{O}$ were not seen then the induced theory $\mathcal{H}_{\mathcal{B}}^{\mathcal{E}\setminus\mathcal{O}}$ would contain the clause $t_p(X) \leftarrow p(X)$ instead of c_2 and c_3, being more concise. In particular, $cov_{\mathcal{E}^+\setminus\mathcal{O}}(\mathcal{H}_{\mathcal{B}}^{\mathcal{E}\setminus\mathcal{O}}) = \frac{\sqrt[3]{5\cdot4\cdot2}}{10}$, $cov_{\mathcal{E}^+\setminus\mathcal{O}}(\mathcal{H}_{\mathcal{B}}^{\mathcal{E}}) = \frac{\sqrt[4]{5\cdot5\cdot2\cdot1}}{10}$ and the gain is 0.08.

Moreover, $\overline{\mathcal{E}} \cup \mathcal{B} \not\rightsquigarrow \overline{\mathcal{O}}$, since if $\{not\_t_p(e)\}$ were not seen then the induced dual theory $\mathcal{H}_{\mathcal{B}}^{\overline{\mathcal{E}}\setminus\overline{\mathcal{O}}}$ would contain the clause $not\_t_p(X) \leftarrow s(X)$ instead of $\overline{c}_1$ and $\overline{c}_2$. In particular, $cov_{\overline{\mathcal{E}}^+}(\mathcal{H}_{\mathcal{B}}^{\overline{\mathcal{E}}\setminus\overline{\mathcal{O}}}) = \frac{\sqrt[3]{5\cdot1\cdot1}}{6}$ and $cov_{\overline{\mathcal{E}}^+}(\mathcal{H}_{\mathcal{B}}^{\overline{\mathcal{E}}}) = \frac{\sqrt[4]{3\cdot1\cdot1\cdot1}}{6}$, and the gain is 0.07.

$\mathcal{O} \subset \mathcal{E}$ pure	$\overline{\mathcal{E}} \cup \mathcal{B} \rightsquigarrow \overline{\mathcal{O}}$	$\overline{\mathcal{E}} \cup \mathcal{B} \not\rightsquigarrow \overline{\mathcal{O}}$
$\mathcal{E} \cup \mathcal{B} \rightsquigarrow \mathcal{O}$	*normal*	*anomalous* (positive) *irregular* (negative)
$\mathcal{E} \cup \mathcal{B} \not\rightsquigarrow \mathcal{O}$	*irregular* (positive) *anomalous* (negative)	*outlier*

Table 1: The different kinds of abnormal examples.

A pure set of examples $\mathcal{O}$ such that both $\mathcal{E} \cup \mathcal{B} \rightsquigarrow \mathcal{O}$ and $\overline{\mathcal{E}} \cup \mathcal{B} \rightsquigarrow \overline{\mathcal{O}}$ 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.

Dually, the same line of reasoning can be adopted to define abnormal examples if a subset $\mathcal{O}$ of negative examples is considered. Table 1 summarizes the different kinds of abnormal example sets that have been defined.

Before concluding we discuss the rationale for defining an abnormality as a set. In some cases two or more related individuals should be removed simultaneously in order to improve generalization of the induced hypothesis. If this happen, then these individuals share some common exceptional properties, which could not be discovered if the instances were considered individually, and then they form all together an abnormal set.

3.4 Statement of the Problem

In this section we define the *outlier detection problem* in the context of Inductive Logic Programming.

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 in the sequel of the section. Intuitively, the novel definition of compliance focuses on the portion of the theory involving the examples in $\mathcal{O}$.

For the sake of simplicity, let $\mathcal{O}$ be a set of positive examples and consider the theory $\mathcal{H}_{\mathcal{B}}^{\mathcal{E}}$ induced in presence of $\mathcal{O}$. The set of clauses in $\mathcal{H}_{\mathcal{B}}^{\mathcal{E}}$ can be partitioned in two groups, that are the clauses in $\mathcal{H}_{\mathcal{B}}^{\mathcal{E}}(\mathcal{O})$ that cover some of the examples in $\mathcal{O}$, and the remaining ones, that are the clauses in $\mathcal{H}_{\mathcal{B}}^{\mathcal{E}} \setminus \mathcal{H}_{\mathcal{B}}^{\mathcal{E}}(\mathcal{O})$.

We call *starting theory* the former one, since it builds on the examples in $\mathcal{O}$, while we note that the latter piece of theory does not directly build 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 $\widehat{\mathcal{O}}$ not in $\mathcal{O}$ and covered only by clauses in $\mathcal{H}_{\mathcal{B}}^{\mathcal{E}}(\mathcal{O})$, and the remaining ones.

Consider now the theory $\mathcal{H}_{\mathcal{B}}^{\mathcal{E} \setminus \mathcal{O}}$ induced in absence of $\mathcal{O}$. In this case, the piece of theory which is directly affected by the absence of the examples in $\mathcal{O}$ is that composed of the clauses that cover some of the examples in $\widehat{\mathcal{O}}$ (recall

that these examples are those not covered by the clauses in $\mathcal{H}^{\mathcal{E}}_{\mathcal{B}} \setminus \mathcal{H}^{\mathcal{E}}_{\mathcal{B}}(\mathcal{O})$). Hence, we name *ending theory* the above described theory.

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

Definition 7 (Compliance) Given a background knowledge $\mathcal{B}$, a set of examples $\mathcal{E}$, and a pure subset $\mathcal{O}$ of $\mathcal{E}$, let $\widehat{\mathcal{E}}$ denote $\mathcal{E}^+ \setminus \mathcal{O}$. We say that $\mathcal{O}$ α-complies (or, simply, complies) with $\mathcal{E} \cup \mathcal{B}$, written $\mathcal{E} \cup \mathcal{B} \rightsquigarrow \mathcal{O}$, if

$$gain_{\widehat{\mathcal{E}}}\left(\overleftarrow{\mathcal{H}}^{\mathcal{E}}_{\mathcal{B}}(\mathcal{O}), \overrightarrow{\mathcal{H}}^{\mathcal{E}}_{\mathcal{B}}(\mathcal{O})\right) < \alpha,$$

holds, where the theories $\overrightarrow{\mathcal{H}}^{\mathcal{E}}_{\mathcal{B}}(\mathcal{O})$ (the *starting theory*) and $\overleftarrow{\mathcal{H}}^{\mathcal{E}}_{\mathcal{B}}(\mathcal{O})$ (the *ending theory*), are defined as follows:

Positive examples: if $\mathcal{O} \subseteq \mathcal{E}^+$, then $\overrightarrow{\mathcal{H}}^{\mathcal{E}}_{\mathcal{B}}(\mathcal{O})$ is $\mathcal{H}^{\mathcal{E}}_{\mathcal{B}}(\mathcal{O})$ and $\overleftarrow{\mathcal{H}}^{\mathcal{E}}_{\mathcal{B}}(\mathcal{O})$ is $\mathcal{H}^{\mathcal{E}\setminus\mathcal{O}}_{\mathcal{B}}(\widehat{\mathcal{O}})$, where $\widehat{\mathcal{O}}$ is

$$covers_{\widehat{\mathcal{E}}}\left(\mathcal{H}^{\mathcal{E}}_{\mathcal{B}}\right) \setminus covers_{\widehat{\mathcal{E}}}\left(\mathcal{H}^{\mathcal{E}}_{\mathcal{B}} \setminus \mathcal{H}^{\mathcal{E}}_{\mathcal{B}}(\mathcal{O})\right),$$

that is the set of examples in $\widehat{\mathcal{E}}$ covered only by clauses of $\mathcal{H}^{\mathcal{E}}_{\mathcal{B}}$ that cover some examples in $\mathcal{O}$;

Negative examples: if $\mathcal{O} \subseteq \mathcal{E}^-$, then $\overleftarrow{\mathcal{H}}^{\mathcal{E}}_{\mathcal{B}}(\mathcal{O})$ is $\mathcal{H}^{\mathcal{E}\setminus\mathcal{O}}_{\mathcal{B}}(\mathcal{O})$ and $\overrightarrow{\mathcal{H}}^{\mathcal{E}}_{\mathcal{B}}(\mathcal{O})$ is $\mathcal{H}^{\mathcal{E}}_{\mathcal{B}}(\widehat{\mathcal{O}})$, where $\widehat{\mathcal{O}}$ is

$$covers_{\widehat{\mathcal{E}}}\left(\mathcal{H}^{\mathcal{E}\setminus\mathcal{O}}_{\mathcal{B}}\right) \setminus covers_{\widehat{\mathcal{E}}}\left(\mathcal{H}^{\mathcal{E}\setminus\mathcal{O}}_{\mathcal{B}} \setminus \mathcal{H}^{\mathcal{E}\setminus\mathcal{O}}_{\mathcal{B}}(\mathcal{O})\right).$$

that is the set of examples in $\widehat{\mathcal{E}} = \mathcal{E}^+$ covered only by clauses of $\mathcal{H}^{\mathcal{E}\setminus\mathcal{O}}_{\mathcal{B}}$ that cover some examples in $\mathcal{O}$.

Next, an example of outlier set follows.

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

$$covers_{\mathcal{E}^+\setminus\{t_p(e)\}}(\{c_1, c_2, c_3, c_4\}) \setminus covers_{\mathcal{E}^+\setminus\{t_p(e)\}}(\{c_1, c_2, c_4\})$$

that is $\{t_p(o)\}$. Let the theory $\mathcal{H}^{\mathcal{E}\setminus\mathcal{O}}_{B}$ induced in absence of $\mathcal{O}$ be

$$\mathcal{H}^{\mathcal{E}\setminus\mathcal{O}}_{\mathcal{B}} = \begin{cases} c'_1 \equiv t_p(X) \leftarrow q(X) & \langle 5\rangle \\ c'_2 \equiv t_p(X) \leftarrow p(X) & \langle 4\rangle \\ c'_3 \equiv t_p(X) \leftarrow t(X) & \langle 2\rangle \end{cases}$$

Then the ending theory $\overleftarrow{\mathcal{H}}^{\mathcal{E}}_{\mathcal{B}}(\mathcal{O})$ is $\mathcal{H}^{\mathcal{E}\setminus\mathcal{O}}_{B}(\widehat{\mathcal{O}}) = \{c'_2\}$, and the gain is $\frac{4}{10} - \frac{1}{10} = 0.30$.

As for the dual concept, $\overline{\mathcal{O}}$, let the theory $\mathcal{H}^{\overline{\mathcal{E}}\setminus\overline{\mathcal{O}}}_{\mathcal{B}}$ induced in absence of $\overline{\mathcal{O}}$ be

$$\mathcal{H}^{\overline{\mathcal{E}}\setminus\overline{\mathcal{O}}}_{\mathcal{B}} = \begin{cases} \overline{c}'_1 \equiv not\_t_p(X) \leftarrow s(X) & \langle 4\rangle \\ \overline{c}'_2 \equiv not\_t_p(j) & \langle 1\rangle \\ \overline{c}'_3 \equiv not\_t_p(k) & \langle 1\rangle \end{cases}$$

Then the ending theory $\overleftarrow{\mathcal{H}}^{\overline{\mathcal{E}}}_{\mathcal{B}}(\overline{\mathcal{O}})$ is $\mathcal{H}^{\overline{\mathcal{E}}\setminus\overline{\mathcal{O}}}_{\mathcal{B}}(\overline{\mathcal{O}}) = \{\overline{c}'_1\}$. The set $\widehat{\overline{\mathcal{O}}}$ is

$$covers_{\mathcal{E}^-}(\{\overline{c}'_1, \overline{c}'_2, \overline{c}'_3\}) \setminus covers_{\mathcal{E}^-}(\{\overline{c}'_2, \overline{c}'_3\}),$$

that is the set consisting of the examples f, n, g, and h. The starting theory $\overrightarrow{H}^{\overline{\mathcal{E}}}_{\mathcal{B}}(\overline{\mathcal{O}})$ is $\mathcal{H}^{\overline{\mathcal{E}}}_{\mathcal{B}}\left(\widehat{\overline{\mathcal{O}}}\right) = \{\overline{c}_1, \overline{c}_2\}$, and the gain is $\frac{4}{6} - \frac{\sqrt{3\cdot 1}}{6} = 0.38$.

Next, it is shown an example of irregular set.

Example 1 (continued). Consider the irregular set $\mathcal{O} = \{t_p(d)\}$. In this case, the starting theory is $\{c_4\}$, $\widehat{\mathcal{O}}$ is empty, and also the ending theory is empty. Then the gain is given by

$$cov_{\mathcal{E}^+\setminus\{t_p(d)\}}(\emptyset) - cov_{\mathcal{E}^+\setminus\{t_p(d)\}}(\{c_4\}),$$

and is equal to $1 - \frac{1}{10} = 0.90$ (we recall that $cov_{\mathcal{E}}(\emptyset) = 1$ by definition).

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 these portions of knowledge are related to 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 $\mathcal{O}$ ($\overline{\mathcal{O}}$, resp.) does not comply with $\mathcal{E} \cup \mathcal{B}$, then we call *direct* (*dual*, resp.) *explanation* the pair $(\overrightarrow{\mathcal{H}}^{\mathcal{E}}_{\mathcal{B}}(\mathcal{O}), \overleftarrow{\mathcal{H}}^{\mathcal{E}}_{\mathcal{B}}(\mathcal{O}))$ $((\overrightarrow{\mathcal{H}}^{\overline{\mathcal{E}}}_{\mathcal{B}}(\overline{\mathcal{O}}), \overleftarrow{\mathcal{H}}^{\overline{\mathcal{E}}}_{\mathcal{B}}(\overline{\mathcal{O}}))$, resp.), also said the *direct* (*dual*, resp.) *starting/ending theories* associated with $\mathcal{O}$ ($\overline{\mathcal{O}}$, resp.) in $\mathcal{H}^{\mathcal{E}}_{\mathcal{B}}$ ($\mathcal{H}^{\overline{\mathcal{E}}}_{\mathcal{B}}$, resp.).

Second, comparing only two 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.

Before providing the formal definition of outlier problem, it is needed to note that abnormal sets should adhere to two additional requirements. First, they should be small, that is the size of the abnormal set should not exceed few units, since outliers are naturally either single instances or at most small groups of instances. Second, the abnormal sets should not contain unnecessary instances. This can be obtained by requiring that these sets are minimal with respect to the property of being abnormal.

Now we are in the position of formally defining the *Concept-Based Outlier Detection* problem.

Definition 8 (Concept-Based Outlier Detection Problem) Given a background knowledge $\mathcal{B}$, a set of examples $\mathcal{E}$, and a maximum size k_{max}, find the minimal irregular, anomalous and outlier subsets $\mathcal{O}$ of $\mathcal{E}$ of size not exceeding k_{max}, together with their associated explanations.

Algorithm: CBOut($\mathcal{B}$, $\mathcal{E}$, α, k_{max})

1: $Out \leftarrow \emptyset$
2: $Anom \leftarrow \emptyset$
3: $Irr \leftarrow \emptyset$
4: find the hypothesis $\mathcal{H}_0 \leftarrow \mathcal{H}^{\mathcal{E}}_{\mathcal{B}}$
5: find the hypothesis $\overline{\mathcal{H}}_0 \leftarrow \mathcal{H}^{\overline{\mathcal{E}}}_{\mathcal{B}}$
6: let $Cand_1 \leftarrow \{\{e\} \mid e \in \mathcal{E}\}$
7: **for** $k = 1$ ***to*** k_{max} **do**
8: let $NextCand_k \leftarrow \emptyset$
9: **foreach** $\mathcal{O}$ ***in*** $Cand_k$ **do**
10: **if** $\mathcal{O} \subseteq \mathcal{E}^+$ **then**
11: let $\langle \mathcal{H}^+_s, \mathcal{H}^+_e, g^+ \rangle \leftarrow Gain^+(\mathcal{H}_0, \mathcal{B}, \mathcal{E}, \mathcal{O})$
12: let $\langle \mathcal{H}^-_s, \mathcal{H}^-_e, g^- \rangle \leftarrow Gain^-(\overline{\mathcal{H}}_0, \mathcal{B}, \overline{\mathcal{E}}, \overline{\mathcal{O}})$
13: **else**
14: let $\langle \mathcal{H}^-_s, \mathcal{H}^-_e, g^- \rangle \leftarrow Gain^-(\mathcal{H}_0, \mathcal{B}, \mathcal{E}, \mathcal{O})$
15: let $\langle \mathcal{H}^+_s, \mathcal{H}^+_e, g^+ \rangle \leftarrow Gain^+(\overline{\mathcal{H}}_0, \mathcal{B}, \overline{\mathcal{E}}, \overline{\mathcal{O}})$
16: **if** $g^+ \geq \alpha$ ***and*** $g^- \geq \alpha$ **then**
17: $Update(Out, \langle \mathcal{O}, (\mathcal{H}^+_s, \mathcal{H}^+_e), (\mathcal{H}^-_s, \mathcal{H}^-_e) \rangle)$
18: **else**
19: **if** $g^+ \geq \alpha$ **then**
20: $Update(Irr, \langle \mathcal{O}, (\mathcal{H}^+_s, \mathcal{H}^+_e) \rangle)$
21: **else if** $g^- \geq \alpha$ **then**
22: $Update(Anom, \langle \mathcal{O}, (\mathcal{H}^-_s, \mathcal{H}^-_e) \rangle)$
23: let $NextCand_k \leftarrow NextCand_k \cup \{\mathcal{O}\}$
24: let $Cand_{k+1} \leftarrow GenerateCand(NextCand_k)$
25: **return** $\langle Out, Irr, Anom \rangle$

Fig. 2: The CBOut algorithm.

4 CBOut Algorithm

In this section we present the CBOut algorithm for mining the Concept-Based outliers. The pseudo-code of the algorithm is reported in Figure 2. It takes as input a background theory $\mathcal{B}$, a set of examples $\mathcal{E}$, and two user-defined parameters, that are α, the gain threshold, and k_{max}, the maximum size of an anomalous set to be found. It outputs three sets, that are, *Out*, containing the outlier sets in $\mathcal{B} \cup \mathcal{E}$, *Irr*, containing the irregular sets in $\mathcal{B} \cup \mathcal{E}$, and *Anom*, containing the anomalous sets in $\mathcal{B} \cup \mathcal{E}$. Each abnormal set is accompanied by its explanation.

First of all, the algorithm finds the hypothesis $\mathcal{H}^{\mathcal{E}}_{\mathcal{B}}$ and the dual hypothesis $\mathcal{H}^{\overline{\mathcal{E}}}_{\mathcal{B}}$ (lines 4-5).

The method visits the pure subsets of examples consisting of at most k_{max} elements in a bottom-up manner. In particular, during the k-th ($k = 1, 2, \ldots, k_{max}$) iteration (lines 7-24) only subsets of examples of size k are considered. The sets taken into account are those stored in the set $Cand_k$, which

Function $Gain^+(\mathcal{H}, \mathcal{B}, \mathcal{E}, \mathcal{O})$

1: let $\mathcal{H}_1 \leftarrow \mathcal{H}(\mathcal{O})$ // *starting theory*
2: find the hypothesis $\mathcal{H}_2 \leftarrow \mathcal{H}_{\mathcal{B}}^{\mathcal{E}\setminus\mathcal{O}}$
3: let $\widehat{\mathcal{E}} \leftarrow \mathcal{E}^+ \setminus \mathcal{O}$
4: let $\widehat{\mathcal{O}} \leftarrow covers_{\widehat{\mathcal{E}}}(\mathcal{H}) \setminus covers_{\widehat{\mathcal{E}}}(\mathcal{H} \setminus \mathcal{H}_1)$
5: let $\mathcal{H}_3 \leftarrow \mathcal{H}_2(\widehat{\mathcal{O}})$ // *ending theory*
6: **return** $\langle \mathcal{H}_1, \mathcal{H}_3, gain_{\widehat{\mathcal{E}}}(\mathcal{H}_3, \mathcal{H}_1)\rangle$

Function $Gain^-(\mathcal{H}, \mathcal{B}, \mathcal{E}, \mathcal{O})$

7: find the hypothesis $\mathcal{H}_2 \leftarrow \mathcal{H}_{\mathcal{B}}^{\mathcal{E}\setminus\mathcal{O}}$
8: let $\mathcal{H}_3 \leftarrow \mathcal{H}_2(\mathcal{O})$ // *ending theory*
9: let $\widehat{\mathcal{E}} \leftarrow \mathcal{E}^+$
10: let $\widehat{\mathcal{O}} \leftarrow covers_{\widehat{\mathcal{E}}}(\mathcal{H}_2) \setminus covers_{\widehat{\mathcal{E}}}(\mathcal{H}_2 \setminus \mathcal{H}_3)$
11: let $\mathcal{H}_1 \leftarrow \mathcal{H}(\widehat{\mathcal{O}})$ // *starting theory*
12: **return** $\langle \mathcal{H}_1, \mathcal{H}_3, gain_{\widehat{\mathcal{E}}}(\mathcal{H}_3, \mathcal{H}_1)\rangle$

Fig. 3: $Gain^+$ and $Gain^-$ functions.

consists in all the potential minimal abnormal sets of size k. The set $Cand_1$ contains $|\mathcal{E}|$ singleton sets $\{e\}$, one for each example e occurring in $\mathcal{E}$ (line 6).

The functions $Gain^+$ and $Gain^-$ (see Figure 3) are used to test compliance according to Definition 7. In particular, $Gain^+$ ($Gain^-$, resp.) takes as input an hypothesis $\mathcal{H}$, a background theory $\mathcal{B}$, a set of examples $\mathcal{E}$, and a subset $\mathcal{O}$ of the positive (negative, resp.) examples in $\mathcal{E}$, and returns the starting and ending theories associated with $\mathcal{O}$ and the associated gain.

For each set $\mathcal{O}$ stored in $Cand_k$, the direct and dual starting and ending theories are computed together with the associated gain values (lines 11-12, for $\mathcal{O}$ a set of positive examples, and lines 14-15, for $\mathcal{O}$ a set of negative examples). In particular, for $\mathcal{O}$ being a set of positive examples, the direct starting $\mathcal{H}_s^+$ and ending $\mathcal{H}_e^+$ theories with the associated the gain g^+ are computed by using the $Gain^+$ function (line 11), while the dual starting $\mathcal{H}_s^-$ and ending $\mathcal{H}_e^-$ theories with the associated gain g^- are computed by using the $Gain^-$ function (line 12). Conversely, for $\mathcal{O}$ being a set of negative examples, the direct starting $\mathcal{H}_s^-$ and ending $\mathcal{H}_e^-$ theories with the associated gain g^- are computed by using the $Gain^-$ function (line 14), while the dual starting $\mathcal{H}_s^+$ and ending $\mathcal{H}_e^+$ theories with the associated gain g^+ are computed by using the $Gain^+$ function (line 15).

Once the gains are computed, it is checked whether the set is abnormal or not. In particular, if g^+ or g^- exceeds α, the set $\mathcal{O}$ is abnormal and it is inserted, together with the associated explanation, in Out, Irr, or $Atom$, on the basis of the kind of abnormality. For minimality, if $\mathcal{O}$ is irregular (anomalous, resp.), it is inserted in Irr ($Anom$, resp.) provided that a subset of $\mathcal{O}$ is not already present in Irr ($Anom$, resp.). Moreover, if $\mathcal{O}$ is not an outlier set,

it is stored in $NextCand_k$ in order to be exploited to generate the potential abnormal sets for the next iteration.

Note that, at the end of the current iteration, $NextCand_k$ contains all the non-outlier sets having size k. In order to build the candidate minimal abnormal sets of size $k+1$, it is needed to find the pure subsets of examples of size $k+1$ such that all their subsets of size k occur in $NextCand_k$. This is taken care of by function *GenerateCand* (line 24).

At the end of the main cycle, the sets Out, Irr, and $Anom$ contain all the minimal outlier, irregular, and anomalous set, respectively, in $\mathcal{B} \cup \mathcal{E}$ of size not greater than k_{max}.

4.1 Computational complexity

In this section we analyze the temporal cost of the CBOut algorithm. We employ the following notation:

- h, denotes the maximum number of clauses in an induced hypothesis; and
- $C_{ind}(\mathcal{B}, \mathcal{E})$, denotes the cost required to induce the hypothesis $\mathcal{H}^{\mathcal{E}}_{\mathcal{B}}$ from the background knowledge $\mathcal{B}$ and the set of examples $\mathcal{E}$.

W.l.o.g., we assume that the inductor outputs, together with the induced hypothesis $\mathcal{H}^{\mathcal{E}}_{\mathcal{B}}$, the sets $covers_{\mathcal{E}}(c)$ of examples covered by each clause c belonging to $\mathcal{H}^{\mathcal{E}}_{\mathcal{B}}$.

First of all, the algorithm induces the hypothesis $\mathcal{H}_0$ and $\overline{\mathcal{H}}_0$ (lines 4-5 of Figure 2), a step which costs $O(C_{ind}(\mathcal{B}, \mathcal{E}))$. Next, the pure subsets of at most $k_{\max}$ elements are considered. The maximum number of pure subsets is $\binom{|\mathcal{E}|}{k_{max}}$ and, then, $O\left(|\mathcal{E}|^{k_{max}}\right)$. For each pure subset, the dominant operations are the computation of the $Gain^+$ and $Gain^-$ functions, whose costs are made explicit next.

In order to evaluate the $Gain^+$ function (see Figure 3), the first step is to compute the starting theory (line 1), which corresponds to find the clauses covering some examples of $\mathcal{O}$. Since the maximum number of clauses is h, the cost required to compute the starting theory is $O(h|\mathcal{E}|)$, corresponding to the cost of determining whether the set of examples covered by a clause has as elements some examples of $\mathcal{O}$.

The next step consists in inducing an hypothesis (line 2) by considering a reduced set of examples, and the associated cost is $C_{ind}(\mathcal{B}, \mathcal{E})$. As for the cost of evaluating the *covers* functions in line 4, it corresponds to the cost of finding the subset of $\widehat{\mathcal{E}}$ covered by some clauses in $\mathcal{H}$ and, then, it requires $O(|\mathcal{E}|)$ elementary operation for each clause of $\mathcal{H}$. Thus, the total cost required to compute the set $\widehat{\mathcal{O}}$ is $O(h|\mathcal{E}|)$.

The cost of computing the ending theory (line 5) is the same of that of computing the starting theory, that is $O(h|\mathcal{E}|)$. As for the last step (line 6), the *gain* function has to be evaluated. On the basis of Equation 1, the cost of computing the *gain* function corresponds to the cost of computing twice the *cov* function. The *cov* function multiplies h factors and, then, its cost is

$O(h)$. Therefore, the total cost required to compute the $Gain^+$ function is $O\big(C_{ind}\,(\mathcal{B},\mathcal{E}) + h|\mathcal{E}|\big)$.

As far as the $Gain^-$ function is concerned, its cost is also $O\big(C_{ind}\,(\mathcal{B},\mathcal{E}) + h|\mathcal{E}|\big)$ and it can be obtained by following a line similar to that employed for the $Gain^+$ function.

Concluding, the cost required by the CBOut algorithm to perform its operations is

$$O\Big(|\mathcal{E}|^{k_{\max}} \cdot \big(C_{ind}(\mathcal{B},\mathcal{E}) + h|\mathcal{E}|\big)\Big).$$

5 Dealing with Complex Domains

When large example sets and background theories are taken into account, the temporal cost of the CBOut algorithm may become heavy. Thus, in order to made the approach here presented feasible on large and real-life domains, in this section we introduce two variants of the basic CBOut algorithm, namely hCBOut and aCBOut. The goal of both methods is to reduce the computational effort while maintaining a good accuracy, but they present different characteristics and are suitable for different scenarios.

The hCBOut variant (for Heuristic CBOut) is designed to alleviate the cost of CBOut in domains where a few hundreds examples are available, possibly together with complex background theories. Intuitively, hCBOut adopts a criterion to select some examples as candidate to form an abnormal set and then explores the power set of this set of candidate examples by using the same strategy of CBOut. hCBOut is consistent (that is, the sets returned are actually abnormal), but not complete (that is, it could not return all the abnormal sets).

The aCBOut variant (for Approximate CBOut) is designed to make feasible the search for abnormal sets in domains where large set of examples are available. To considerably improve efficiency, aCBOut adopts a criterion to directly select some candidate abnormal set of examples and then checks for their abnormality. Similarly to hCBOut, aCBOut is consistent and incomplete. Morever, its solutions could not be minimal with respect to the containment between abnormal sets.

The rest of the section presents in detail the hCBOut (Section 5.1) and aCBOut (Section 5.2) algorithms.

5.1 Heuristic CBOut algorithm

Basically, the hCBOut algorithm (for Heuristic CBOut) is based on the selection of a subset $\mathcal{E}_{Cand}$ of the input set of examples $\mathcal{E}$ to be used as *candidate* abnormal examples by the outlier detector algorithm, with $|\mathcal{E}_{Cand}| \ll |\mathcal{E}|$, which allows to reduce the overall cost of the mining algorithm to

$$O\Big(|\mathcal{E}_{Cand}|^{k_{\max}} \cdot \big(C_{ind}(\mathcal{B},\mathcal{E}) + h|\mathcal{E}|\big)\Big).$$

The selection of the candidate examples is accomplished by exploiting suitable scoring functions introduced next. Informally speaking, the scoring functions assign to each example a score aiming at reflecting its propension to represent an irregular or an anomalous fact. Examples are then ranked on the basis of these scoring functions, and the top ranked examples are collected to form the set $\mathcal{E}_{Cand}$ of the candidate abnormal examples.

First, Sections 5.1.1, 5.1.2, and 5.1.3 introduce, respectively, some preliminary definitions, the irregularity, and the anomaly score functions and, then, subsequent Sections 5.1.4 and 5.1.5 describe how these scores can be exploited in order to build the set of candidate abnormal examples and the hCBOut algorithm.

5.1.1 Similarity between clauses and score function

Some preliminary definitions are needed in order to provide the irregularity and the anomaly score. Let C be a set of clauses and E be a set of examples. The set $edges_C(e)$ of *edges* of e in C, is composed of the clauses of C that cover at least e, that is to say

$$edges_C(e) = \{c \in C : e \in covers_{\mathcal{E}}(c)\},$$

while the set $neighs_{C,E}(e)$ of *neighbors* of e in E according to C, is composed of the examples of E that share at least an edge in C with e, that is:

$$neighs_{C,E}(e) = \{e' \in E : edges_C(e) \cap edges_C(e') \neq \emptyset\}.$$

Let c and c' be two clauses. Then, the *similarity* $sim_E(c, c')$ between c and c' w.r.t. the set of examples E is defined as

$$sim_E(c, c') = \frac{|covers_E(c) \cap covers_E(c')|}{1 + |covers_E(c) \,\Delta\, covers_E(c')|},$$

where Δ denotes the symmetric difference set theoretic binary operator, that is $A \,\Delta\, B = (A \cup B) \setminus (A \cap B)$.

Specifically, the larger the number of examples in E common to c and c' and the larger their similarity and, the smaller the number of examples covered by either c or c', but not by both of them, and the larger their similarity.

Let e be a positive (negative, resp.) example, then $pos_e(\mathcal{E})$ denotes the set $\mathcal{E}$ ($\overline{\mathcal{E}}$, resp.). Moreover, $neg_e(\mathcal{E})$ denotes the set $\overline{pos_e(\mathcal{E})}$, that is $\overline{\mathcal{E}}$ ($\mathcal{E}$, resp.) for e a positive (negative, resp.) example. The set $neg_e(\mathcal{E})^+$ is said to be the set of the *opposite examples* of e.

Given an example e, two sets of clauses C_0 and C_1, and a set of examples E, the function $score_{C_0,C_1,E}(e)$ is defined as:

$$score_{C_0,C_1,E}(e) = \sum_{\left(c \in edges_{C_1}(e)\right)} \sum_{\left(e' \in neighs_{C_1,E}(e)\right)} \sum_{\left(c' \in edges_{C_0}(e')\right)} \frac{|covers_E(c) \cap covers_E(c')|}{|covers_E(c) \,\Delta\, covers_E(c')| + 1}.$$

The above defined function is used in the following in order to define the irregularity and anomaly score functions. Intuitions underlying this score are discussed next.

Here, the computational effort of evaluating the score function is analyzed. First of all, the cost of the functions therein employed is depicted. The cost of computing the set $edges_C(e)$ is linear in $|C|$, since for each clause $c \in C$ it must be checked if e belongs to the set of examples covered by c. As for the $neighs_{C,E}(e)$ set, the cost of computing it corresponds to the cost of computing the union of the sets of examples covered by the clauses in $edges_C(e)$ and then it is $O(|edges_C(e)| \cdot |E|) = O(|C| \cdot |E|)$. The cost of the similarity $sim_E(c, c')$ between two clauses c and c' corresponds to the cost of computing the intersection and the symmetric difference of two sets of examples. Since the size of each set is $O(|E|)$, the similarity function requires $O(|E|)$ time.

Notice that the size of the $edges_C(e)$ sets is at most equal to the number $|C|$ of clauses and the size of the $neighs_{C,E}(e)$ set is at most equal to the number of examples $|E|$. Thus, as far as the score function is concerned, in order to evaluate it a computational effort of $O(|C_1|\cdot|E|+|C_0|\cdot|E|+|C_1|\cdot|E|\cdot|C_0|\cdot|E|) = O(|C_1| \cdot |C_0| \cdot |E|^2)$ is required.

5.1.2 Irregularity score function

The irregularity score function $score_{ir}(e)$, next introduced, assigns to the example e a score (a positive rational number) aiming to reflect its propension to become part of an irregular set.

We are now in the position of defining the *irregularity score* $score_{ir}$. Let

$$\mathcal{E}_p = pos_e(\mathcal{E}), \text{ and } \mathcal{H}_0 = \mathcal{H}_{\mathcal{B}}^{\mathcal{E}_p},$$

then:

$$score_{ir}(e) = score_{\mathcal{H}_0,\mathcal{H}_0,\mathcal{E}_p^+}(e),$$

that is:

$$score_{ir}(e) = \sum_{\left(c \in edges_{\mathcal{H}_0}(e)\right)} \sum_{\left(e' \in neighs_{\mathcal{H}_0,\mathcal{E}_p^+}(e)\right)} \sum_{\left(c' \in edges_{\mathcal{H}_0}(e')\right)} sim_{\mathcal{E}_p^+}(c, c').$$

Promising irregular examples e are those having associated a small value $score_{ir}(e)$.

Intuitively, irregular examples share few commonalities with the examples coming from the same concept, and the above summation aims at providing a way to measure this intuition.

Indeed, since the clauses belonging to the induced hypothesis $\mathcal{H}_{\mathcal{B}}^{\mathcal{E}}$ capture regularities in the example set, the irregularity score relates the similarity between two examples to the similarity among the regularities in which they are involved.

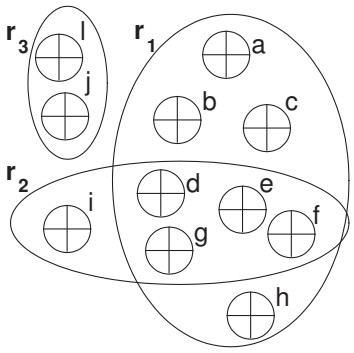


Fig. 4: Example for the irregularity score.

Example 2 Figure 4 reports the hypothesis $\mathcal{H}^{\mathcal{E}}_{\mathcal{B}} = \{r_1, r_2, r_3\}$ induced on the set of positive examples $\mathcal{E} = \{a, b, c, d, e, f, g, h, i, j, l\}$. Consider the examples i, j and d. Intuitively, the example d is not likely to be irregular, since it shares commonalities with many examples and then it is expected to have a large value of irregularity score. Conversely, example j shares commonalities only with the example l and then it is likely to have a very small value of irregularity score.

First of all, the similarity between the pairs of clauses belonging to $\mathcal{H}^{\mathcal{E}}_{\mathcal{B}}$ is:

$$sim_{\mathcal{E}}(r_1, r_1) = \frac{8}{1}, \qquad sim_{\mathcal{E}}(r_1, r_2) = \frac{4}{6}, \qquad sim_{\mathcal{E}}(r_1, r_3) = \frac{0}{11},$$
$$sim_{\mathcal{E}}(r_2, r_2) = \frac{5}{1}, \qquad sim_{\mathcal{E}}(r_2, r_3) = \frac{0}{8}, \qquad sim_{\mathcal{E}}(r_3, r_3) = \frac{2}{1}.$$

Moreover, note that

$$edges_{\mathcal{H}^{\mathcal{E}}_{\mathcal{B}}}(a) = edges_{\mathcal{H}^{\mathcal{E}}_{\mathcal{B}}}(b) = edges_{\mathcal{H}^{\mathcal{E}}_{\mathcal{B}}}(c) = edges_{\mathcal{H}^{\mathcal{E}}_{\mathcal{B}}}(h), \quad \text{and}$$
$$edges_{\mathcal{H}^{\mathcal{E}}_{\mathcal{B}}}(d) = edges_{\mathcal{H}^{\mathcal{E}}_{\mathcal{B}}}(g) = edges_{\mathcal{H}^{\mathcal{E}}_{\mathcal{B}}}(e) = edges_{\mathcal{H}^{\mathcal{E}}_{\mathcal{B}}}(f); \quad \text{then}$$
$$neighs_{\mathcal{H}^{\mathcal{E}}_{\mathcal{B}},\mathcal{E}}(a) = neighs_{\mathcal{H}^{\mathcal{E}}_{\mathcal{B}},\mathcal{E}}(b) = neighs_{\mathcal{H}^{\mathcal{E}}_{\mathcal{B}},\mathcal{E}}(c) = neighs_{\mathcal{H}^{\mathcal{E}}_{\mathcal{B}},\mathcal{E}}(h), \quad \text{and}$$
$$neighs_{\mathcal{H}^{\mathcal{E}}_{\mathcal{B}},\mathcal{E}}(d) = neighs_{\mathcal{H}^{\mathcal{E}}_{\mathcal{B}},\mathcal{E}}(g) = neighs_{\mathcal{H}^{\mathcal{E}}_{\mathcal{B}},\mathcal{E}}(e) = neighs_{\mathcal{H}^{\mathcal{E}}_{\mathcal{B}},\mathcal{E}}(f).$$

As far as the example d is concerned, it holds that:

$$edges_{\mathcal{H}^{\mathcal{E}}_{\mathcal{B}}}(d) = \{r_1, r_2\}, \qquad neighs_{\mathcal{H}^{\mathcal{E}}_{\mathcal{B}},\mathcal{E}}(d) = \{a, b, c, d, e, f, g, h, i\}, \text{ and}$$

$$score_{ir}(d) = \sum_{c \in \{r_1, r_2\}} \left(\sum_{e' \in \{a,b,c,h\}} \sum_{c' \in \{r_1\}} sim_{\mathcal{E}}(c, c') \quad + \right.$$
$$\left. + \sum_{e' \in \{d,e,f,g\}} \sum_{c' \in \{r_1, r_2\}} sim_{\mathcal{E}}(c, c') + \sum_{e' \in \{i\}} \sum_{c' \in \{r_2\}} sim_{\mathcal{E}}(c, c') \right)$$
$$= \sum_{c \in \{r_1, r_2\}} \left(4 \cdot sim_{\mathcal{E}}(c, r_1) + 4 \cdot \sum_{c' \in \{r_1, r_2\}} sim_{\mathcal{E}}(c, c') + sim_{\mathcal{E}}(c, r_2) \right) =$$
$$\approx \quad 97.67.$$

Consider, now, the example i. It holds that:

$$edges_{\mathcal{H}^{\mathcal{E}}_{\mathcal{B}}}(i) = \{r_2\}, \qquad neighs_{\mathcal{H}^{\mathcal{E}}_{\mathcal{B}},\mathcal{E}}(i) = \{d,e,f,g,i\}, \text{ and}$$

$$score_{ir}(i) = \sum_{c\in\{r_2\}} \left(\sum_{e'\in\{d,e,f,g\}} \sum_{c'\in\{r_1,r_2\}} sim_{\mathcal{E}}(c,c') \quad + \sum_{e'\in\{i\}} \sum_{c'\in\{r_2\}} sim_{\mathcal{E}}(c,c') \right) =$$
$$= \quad 4 \cdot \sum_{c'\in\{r_1,r_2\}} sim_{\mathcal{E}}(r_2,c') + sim_{\mathcal{E}}(r_2,r_2) \approx 27.67.$$

Finally, for the example j it holds that:

$$edges_{\mathcal{H}^{\mathcal{E}}_{\mathcal{B}}}(j) = \{r_3\}, \qquad neighs_{\mathcal{H}^{\mathcal{E}}_{\mathcal{B}},\mathcal{E}}(j) = \{j,l\}, \text{ and}$$
$$score_{ir}(j) = \sum_{c\in\{r_3\}} \sum_{e'\in\{j,l\}} \sum_{c'\in\{r_3\}} sim_{\mathcal{E}}(c,c') = 2 \cdot sim_{\mathcal{E}}(r_3,r_3) = 4.$$

5.1.3 Anomaly score function

The anomaly score function, next introduced, assigns to the example e a score (a positive rational number) aiming to reflect its propension to become part of an anomalous set.

In order to define the anomaly score, we need to relax the *consistency* property of an induced hypothesis. In particular, given a non-negative integer ϱ, a background theory $\mathcal{B}$, and a set of examples $\mathcal{E}$, we say that $\widetilde{\mathcal{H}}^{\mathcal{E}}_{\mathcal{B},\varrho}$ is a *ϱ-consistent hypothesis* on $\mathcal{E}$ w.r.t. $\mathcal{B}$, if it is complete and, moreover, for each $c \in \widetilde{\mathcal{H}}^{\mathcal{E}}_{\mathcal{B},\varrho}$, it holds that $|covers(c) \cap \mathcal{E}^-| \leq \varrho$. In the sequel, for the sake of readability, we will denote a ϱ-consistent hypothesis $\mathcal{H}^{\mathcal{E}}_{\mathcal{B},\varrho}$ simply as $\mathcal{H}^{\mathcal{E}}_{\varrho}$, thus omitting the specification of the background knowledge $\mathcal{B}$ as subscript. Clearly, a 0-consistent hypothesis is both complete and consistent. An example $e \in \mathcal{E}^-$ such that $\widetilde{\mathcal{H}}^{\mathcal{E}}_{\varrho} \models e$ is said to be *misclassified* in $\widetilde{\mathcal{H}}^{\mathcal{E}}_{\varrho}$.

We are now in the position of defining the *anomaly score* $score^{\varrho}_{an}$. Let

$$\mathcal{E}_n = neg_e(\mathcal{E}),\ \mathcal{H}_0 = \widetilde{\mathcal{H}}^{\mathcal{E}_n}_{\varrho}, \text{ and } \mathcal{H}_1 = \widetilde{\mathcal{H}}^{\mathcal{E}_n}_{\varrho+1},$$

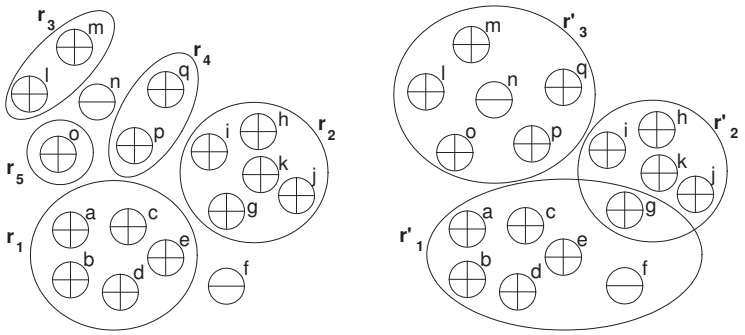
then

$$score^{\varrho}_{an}(e) = \begin{cases} score_{\mathcal{H}_1,\mathcal{H}_0,\mathcal{E}_n}(e) & \text{, if } edges_{\mathcal{H}_1}(e) \neq \emptyset \\ +\infty & \text{, otherwise} \end{cases}$$

where:

$$score_{\mathcal{H}_1,\mathcal{H}_0,\mathcal{E}_n}(e) =$$
$$\sum_{\left(c\in edges_{\mathcal{H}_1}(e)\right)} \sum_{\left(e'\in neighs_{\mathcal{H}_1,\mathcal{E}^+_n}(e)\right)} \sum_{\left(c'\in edges_{\mathcal{H}_0}(e')\right)} sim_{\mathcal{E}^+_n}(c,c').$$

The anomaly score computes the similarity, with respect to the set $neg_e(\mathcal{E})^+$ of the opposite examples of e, among the clauses c that cover e and belong to an inconsistent hypothesis $\mathcal{H}_1$ and the clauses c' that cover the opposite neighbors of e according to $\mathcal{H}_1$ and belong to its preceding hypothesis $\mathcal{H}_0$.



(a) Consistent hypothesis. (b) 1-consistent hypothesis.

Fig. 5: Example for the anomaly score.

Recall that $\widetilde{\mathcal{H}}^{\mathcal{E}_n}_{\varrho+1}$ is an inconsistent hypothesis on $neg_e(\mathcal{E})$. Thus, the set of clauses $edges_{\widetilde{\mathcal{H}}^{\mathcal{E}_n}_{\varrho+1}}(e)$ is non-empty if e is misclassified in $\widetilde{\mathcal{H}}^{\mathcal{E}_n}_{\varrho+1}$, and it is empty otherwise. In the latter case the score evaluates to $+\infty$, while in the former one it is finite and strictly positive.

In particular, the more similar the two hypotheses, the larger the value of the anomaly score, and the smaller the advantage of having the example as a false negative. Vice versa, a small value for the anomaly score denotes the presence in the inconsistent hypothesis of clauses much more general than those already included in its preceding hypothesis. Since, intuitively, anomalous examples are those preventing the induction of clauses covering much more examples than those actually covered, the examples e having associated a small value $score^{\varrho}_{an}(e)$ can be considered promising anomalous examples.

Example 3 Figures 5a and 5b report the consistent and 1-consistent hypotheses induced on a set of examples. Consider the negative examples f and n. Both these negative examples are covered in the 1-consistent hypothesis. Intuitively, the presence of n breaks a regularity among many examples, while the presence of f avoids just one example, g, to be included in the regularity including examples a, b, c, d and e. Thus, it is expected that the irregularity score of n is much greater than that of f.
Consider, first, the negative example f. It holds that:

$$edges_{\widetilde{\mathcal{H}}_1}(f) = \{r'_1\}, \qquad neighs_{\widetilde{\mathcal{H}}_1, neg_f(\mathcal{E}^+)}(f) = \{a,b,c,d,e,g\}, \text{ and}$$

$$score^0_{an}(f) = \sum_{c\in\{r'_1\}} \left(\sum_{e'\in\{a,b,c,d,e\}} \sum_{c'\in\{r_1\}} sim(c,c') + \sum_{e'\in\{g\}} \sum_{c'\in\{r_2\}} sim(c,c') \right) =$$

$$= \; 5 \cdot sim(r'_1, r_1) + sim(r'_1, r_2) = 5 \cdot \frac{5}{2} + \frac{1}{10} \approx 12.6.$$

Consider, now, the negative example n. It holds that:

$$edges_{\widetilde{\mathcal{H}}_1}(n) = \{r'_3\}, \qquad neighs_{\widetilde{\mathcal{H}}_{\varrho+1}, neg_n(\mathcal{E}^+)}(n) = \{l, m, o, p, q\}, \text{ and}$$

$$score^0_{an}(n) = \sum_{c \in \{r'_3\}} \left(\sum_{e' \in \{l,m\}} \sum_{c' \in \{r_3\}} sim(c, c') + \sum_{e' \in \{o\}} \sum_{c' \in \{r_5\}} sim(c, c') + \right.$$

$$\left. + \sum_{e' \in \{p,q\}} \sum_{c' \in \{r_4\}} sim(c, c') \right) =$$

$$= \quad 2 \cdot sim(r'_3, r_3) + sim(r'_3, r_5) + 2 \cdot sim(r'_3, r_4) = 2 \cdot \frac{2}{4} + \frac{1}{5} + 2 \cdot \frac{2}{4} \approx 2.2.$$

5.1.4 Computing the candidate abnormal examples

In this section we show how the above introduced scores can be employed in order to determine the set $\mathcal{E}_{Cands}$ of candidate abnormal examples.

Figure 6 shows the hCBOut-*candidate-selector* algorithm, that given a background knowledge $\mathcal{B}$, a set of examples $\mathcal{E}$, and the two positive integers m and k_{max}, determines the set $\mathcal{E}_{Cands}$. In particular, the parameter m is used to control the size of the set $\mathcal{E}_{Cands}$.

The algorithm determines four sets of candidate abnormal examples, each of size m. The sets are: candidate positive irregulars (lines 2-5), candidate negative irregulars (lines 6-9), candidate positive anomalous (lines 10-19), and candidate negative anomalous (lines 20-29).

In order to determine these candidate examples, a score is assigned to each example by exploiting the score functions introduced in the preceding sections, that is the irregularity score function for selecting the candidate positive and negative irregular examples (see Section 5.1.2), and the anomaly score for selecting the candidate positive and negative anomalous examples (see Section 5.1.3).

Specifically, ϱ-consistent hypotheses with ϱ varying from 1 to k_{max} are considered to single out candidate anomalous examples. This is accomplished in order to detect potential anomalous sets of size up to k_{max}. We note that a ϱ-consistent hypothesis can be computed by allowing the inductor to induce clauses covering at most ϱ negative examples.

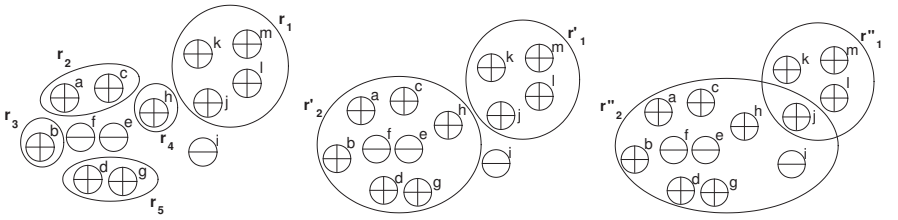
The set $\mathcal{E}_{Cand}$ returned by the algorithm is the union of the four sets above listed.

Here we argument why the ϱ-consistent hypothesis is compared with the preceding one (the $(\varrho-1)$-consistent one) and not with the consistent hypothesis. The motivation is that while the former strategy allows to detect genuine candidate anomalous examples, the latter strategy may lead to the individuation of false candidates. Clearly, the larger the value of ϱ, the more general the ϱ-consistent hypothesis $\widetilde{\mathcal{H}}^{\mathcal{E}_n}_{\varrho}$ and, consequently, the smaller the value of the score $score_{\widetilde{\mathcal{H}}^{\mathcal{E}_n}_{\varrho}, \widetilde{\mathcal{H}}^{\mathcal{E}_n}_{0}, \mathcal{E}_n}(e)$ obtained by comparing the hypothesis $\widetilde{\mathcal{H}}^{\mathcal{E}_n}_{\varrho}$ with the consistent one $\widetilde{\mathcal{H}}^{\mathcal{E}_n}_{0} = \mathcal{H}^{\mathcal{E}_n}_{\mathcal{B}}$. Conversely, the score $score_{\widetilde{\mathcal{H}}^{\mathcal{E}_n}_{\varrho}, \widetilde{\mathcal{H}}^{\mathcal{E}_n}_{\varrho-1}, \mathcal{E}_n}(e)$ obtained by comparing the ϱ-consistent hypothesis $\widetilde{\mathcal{H}}^{\mathcal{E}_n}_{\varrho}$ with the $(\varrho-1)$-consistent one

Function *hCBOut-candidate-selector*($\mathcal{B}$, $\mathcal{E}$, m, k_{max})

set $\mathcal{E}_{Cands}$ to $\emptyset$
// Compute candidate positive irregulars
induce the hypothesis $\mathcal{H}_0^p = \mathcal{H}_{\mathcal{B}}^{\mathcal{E}}$
foreach $e \in \mathcal{E}^+$ **do**
    compute the irregularity score $is_e^+ = score_{\mathcal{H}_0^p, \mathcal{H}_0^p, \mathcal{E}^+}(e)$
insert into $\mathcal{E}_{Cands}$ the examples e associated with the m smallest scores is_e^+
// Compute candidate negative irregulars
induce the hypothesis $\mathcal{H}_0^n = \mathcal{H}_{\mathcal{B}}^{\overline{\mathcal{E}}}$
foreach $e \in \mathcal{E}^-$ **do**
    compute the irregularity score $is_e^- = score_{\mathcal{H}_0^n, \mathcal{H}_0^n, \overline{\mathcal{E}}^+}(e)$
insert into $\mathcal{E}_{Cands}$ the examples e associated with the m smallest scores is_e^-
// Compute candidate positive anomalous
foreach $e \in \mathcal{E}^+$ **do**
    set as_e^+ to $+\infty$
set $\widetilde{\mathcal{H}}_0^n$ to $\mathcal{H}_0^n$
for $\varrho = 1$ ***to*** $k_{\max}$ **do**
    induce the ϱ-consistent hypothesis $\widetilde{\mathcal{H}}_\varrho^n$ on $\overline{\mathcal{E}}$ w.r.t. $\mathcal{B}$
    foreach $e \in \mathcal{E}^+$ **do**
        if $edges_{\widetilde{\mathcal{H}}_\varrho^n}(e) \neq \emptyset$ **then**
            compute the anomaly score $as_{cur}^+ = score_{\widetilde{\mathcal{H}}_\varrho^n, \widetilde{\mathcal{H}}_{\varrho-1}^n, \overline{\mathcal{E}}^+}(e)$
            set as_e^+ to $\min\{as_e^+, as_{cur}^+\}$
insert into $\mathcal{E}_{Cands}$ the examples e associated with the m smallest scores as_e^+ among those smaller than $+\infty$
// Compute candidate negative anomalous
foreach $e \in \mathcal{E}^-$ **do**
    set as_e^- to $+\infty$
set $\widetilde{\mathcal{H}}_0^p$ to $\mathcal{H}_0^p$
for $\varrho = 1$ ***to*** $k_{\max}$ **do**
    induce the ϱ-consistent hypothesis $\widetilde{\mathcal{H}}_\varrho^p$ on $\mathcal{E}$ w.r.t. $\mathcal{B}$
    foreach $e \in \mathcal{E}^-$ **do**
        if $edges_{\widetilde{\mathcal{H}}_\varrho^p}(e) \neq \emptyset$ **then**
            compute the anomaly score $as_{cur}^- = score_{\widetilde{\mathcal{H}}_\varrho^p, \widetilde{\mathcal{H}}_{\varrho-1}^p, \mathcal{E}^+}(e)$
            set as_e^- to $\min\{as_e^-, as_{cur}^-\}$
insert into $\mathcal{E}_{Cands}$ the examples e associated with the m smallest scores as_e^- among those smaller than $+\infty$
return $\mathcal{E}_{Cands}$

Fig. 6: *h*CBOut-*candidate-selector* function



(a) $\varrho \in \{0, 1\}$ (b) $\varrho = 2$ (c) $\varrho = 3$

Fig. 7: Example of anomalous scores.

$\widetilde{\mathcal{H}}^{\mathcal{E}_n}_{\varrho-1}$ is low if e belongs to a set of size ϱ representing a minimal anomalous set, namely such that all its subsets of size $\varrho-1$ are not very likely to represent anomalous sets.

Example 4 As an example, consider the set of examples reported in Figure 7. Figure 7b shows the 2-consistent hypothesis. By comparing this hypothesis with that 1-consistent one reported in Figure 7a, we can see that in the former hypothesis there is a rule, r'_2, that covers a large set of positive examples and generalizes rules r_2, r_3, r_4, and r_5. Furthermore, rule r'_2 covers also the two negative examples e and f. Figure 7c shows the 3-consistent hypothesis. Now rule r''_2 covers a larger set of positive examples than r'_2 and, moreover, also the negative example j.

By comparing the various hypotheses, it is clear that it is advantageous to consider a 2-consistent hypothesis since in this case four rules are generalized in a single one, while the 3-consistent hypothesis offers a limited advantage if compared to the 2-consistent one, but appears to offer about the same advantages of the 2-consistent if compared to the consistent one.

As for the anomaly scores, the value of the scores $score^2_{an}(e)$ and $score^2_{an}(f)$ is approximately to 1.93, while the scores $score^3_{an}(j)$, $score^3_{an}(e)$ and $score^3_{an}(f)$ evaluate to 18.1. Conversely, if the 3-consistent hypothesis is compared with the consistent one it holds that $score_{\widetilde{\mathcal{H}}^{\mathcal{E}_n}_3,\widetilde{\mathcal{H}}^{\mathcal{E}_n}_0,\mathcal{E}_n}(e)$, $score_{\widetilde{\mathcal{H}}^{\mathcal{E}_n}_3,\widetilde{\mathcal{H}}^{\mathcal{E}_n}_0,\mathcal{E}_n}(f)$ and $score_{\widetilde{\mathcal{H}}^{\mathcal{E}_n}_3,\widetilde{\mathcal{H}}^{\mathcal{E}_n}_0,\mathcal{E}_n}(j)$ evaluate about to 1.24. Thus, according to the latter score, the example j would be a good candidate anomalous example.

As for the cost of the hCBOut-*candidate-selector* function, first the irregularity scores for each positive example (lines 3-4) and for each negative example (lines 7-8) have to be computed. These steps cost $O\left(|\mathcal{H}^p_0|^2 \cdot |\mathcal{E}^+|^3\right)$ and $O\left(|\mathcal{H}^n_0|^2 \cdot |\mathcal{E}^-|^3\right)$, respectively (see Section 5.1.1 for the cost of computing the scores). Next, the anomaly scores for the positive (negative, resp.) examples have to be evaluated. This step needs to induce $O(k_{\max})$ inconsistent hypothesis, line 14 (line 24, resp.), and then to compute the anomaly score. Let $C_{ind}(\mathcal{B}, \mathcal{E})$ be the cost of inducing an hypothesis and let h be the maximum number of clauses in an induced hypothesis, then the cost of the hCBOut-*candidate-selector* function is $O\left(k_{max} \cdot \left(C_{ind}(\mathcal{B}, \mathcal{E}) + h^2|\mathcal{E}|^3\right)\right)$.

Function *hCBOut*($\mathcal{B}$, $\mathcal{E}$, α, m, k_{max})

// First phase: candidate selection
1: $\mathcal{E}_{Cand} \leftarrow$ hCBOut-*candidate-selector*($\mathcal{B}$, $\mathcal{E}$, m, k_{max})
// Second phase: search space visit
2: $\langle Out, Irr, Anom\rangle \leftarrow$ CBOut($\mathcal{B}$, $\mathcal{E}$, $\mathcal{E}_{Cand}$, α, k_{max})
3: **return** $\langle Out, Irr, Anom\rangle$

Fig. 8: The hCBOut algorithm.

5.1.5 The hCBOut Algorithm

Figure 8 reports the Heuristic CBOut algorithm. It consists of two phases. The first phase is that of *candidate selection*, which is accomplished by executing the function hCBOut-*candidate-selector* described in Section 5.1.4. This function returns the set $\mathcal{E}_{Cand}$ of abnormal candidates. During the second phase, the abnormal sets are computed through the CBOut algorithm, described in Section 4, by considering as search space only the examples in the set $\mathcal{E}_{Cand}$. The pseudo-code of the CBOut algorithm (see Figure 2) has to be slightly modified in order to limit the search space to the set $\mathcal{E}_{Cand}$. In particular, the set $\mathcal{E}$ employed in line 6 of Figure 2 has to be replaced by the set $\mathcal{E}_{Cand}$, so that the set $Cand_1$ is set to $\{\{e\} \mid e \in \mathcal{E}_{Cand}\}$.

Next, the cost of the hCBOut algorithm is analyzed. The algorithm requires the evaluation of the hCBOut-*candidate-selector* function and the execution of the algorithm CBOut, for a total cost of

$$O\left(k_{max} \cdot \left(C_{ind}(\mathcal{B},\mathcal{E}) + h^2|\mathcal{E}|^3\right) + |\mathcal{E}_{Cands}|^{k_{\max}} \cdot \left(C_{ind}(\mathcal{B},\mathcal{E}) + h|\mathcal{E}|\right)\right),$$

which can be approximated to the second term due the presence of the term $|\mathcal{E}_{Cands}|^{k_{max}}$, that is to

$$O\left(|\mathcal{E}_{cands}|^{k_{\max}} \cdot \left(C_{ind}(\mathcal{B},\mathcal{E}) + h|\mathcal{E}|\right)\right).$$

Before concluding, we briefly discuss on the rationale underlying the introduction of the irregularity and anomaly scores for selecting candidate examples. As a matter of fact, one could argue that the compliance relationship could be directly exploited in order to detect promising abnormal examples, by measuring the gain involving two suitably related hypotheses. However, we point out that the compliance intends to measure the difference of generalization between the hypothesis induced in absence of a certain set of examples and the hypothesis induced on the whole example set. Thus, in order to be meaningfully applied it is needed to precisely induce the hypothesis under the assumption that the set of examples of interest is not seen. But, this is exactly what we want to avoid: that is to say, to enumerate all the subsets of the set of examples and to measure the associated gain value.

Function *aCBOut-candidate-selector*($\mathcal{B}$, $\mathcal{E}$, m, k_{max})

// Compute candidate negative abnormal examples
set $\mathcal{O}^n_{Cands}$ to $\emptyset$
induce the hypothesis $\widetilde{\mathcal{H}}^p_0 = \mathcal{H}^{\mathcal{E}}_{\mathcal{B}}$
for $\varrho = 1$ ***to*** $k_{\max}$ **do**
  induce the ϱ-consistent hypothesis $\widetilde{\mathcal{H}}^p_\varrho$ on $\mathcal{E}$ w.r.t. $\mathcal{B}$
  foreach $c \in \widetilde{\mathcal{H}}^p_\varrho$ **do**
    if $covers_{\mathcal{E}^-}(c) \neq \emptyset$ **then**
      set scr to 1
      foreach $c' \in \widetilde{\mathcal{H}}^p_{\varrho-1}$ **do**
        $scr = \frac{\max\left(scr, |covers_{\mathcal{E}^+}(c) \setminus covers_{\mathcal{E}^+}(c')| + 1\right)}{\frac{1}{|\widetilde{\mathcal{H}}^p_\varrho| - 1} \sum_{c'' \in \widetilde{\mathcal{H}}^p_\varrho \setminus \{c\}} |covers_{\mathcal{E}^+}(c'')|}$
      update the set $\mathcal{O}^n_{Cands}$ by adding the set $covers_{\mathcal{E}^-}(c)$ with associated score scr
leave in $\mathcal{O}^n_{Cands}$ only the sets having associated the m largest scores
// Compute candidate positive abnormal examples
set $\mathcal{O}^p_{Cands}$ to $\emptyset$
induce the hypothesis $\widetilde{\mathcal{H}}^n_0 = \mathcal{H}^{\overline{\mathcal{E}}}_{\mathcal{B}}$
for $\varrho = 1$ ***to*** $k_{\max}$ **do**
  induce the ϱ-consistent hypothesis $\widetilde{\mathcal{H}}^n_\varrho$ on $\overline{\mathcal{E}}$ w.r.t. $\mathcal{B}$
  foreach $c \in \widetilde{\mathcal{H}}^n_\varrho$ **do**
    if $covers_{\mathcal{E}^+}(c) \neq \emptyset$ **then**
      set scr to 1
      foreach $c' \in \widetilde{\mathcal{H}}^n_{\varrho-1}$ **do**
        $scr = \frac{\max\left(scr, |covers_{\mathcal{E}^-}(c) \setminus covers_{\mathcal{E}^-}(c')| + 1\right)}{\frac{1}{|\widetilde{\mathcal{H}}^n_\varrho| - 1} \sum_{c'' \in \widetilde{\mathcal{H}}^n_\varrho \setminus \{c\}} |covers_{\mathcal{E}^-}(c'')|}$
      update the set $\mathcal{O}^p_{Cands}$ by adding the set $covers_{\mathcal{E}^+}(c)$ with associated score scr
leave in $\mathcal{O}^p_{Cands}$ only the sets having associated the m largest scores
return $\mathcal{O}^p_{Cands} \cup \mathcal{O}^n_{Cands}$

Fig. 9: *aCBOut-candidate-selector* function.

5.2 The Approximate CBOut algorithm

Rather than limiting the search space to the subsets of size up to k_{max} of a selected subset $\mathcal{E}_{Cands}$ of the whole set of examples, as done by the hCBOut algorithm, the Approximate CBOut algorithm (aCBOut, for short) reduces further the size of the search space by directly selecting a small set $\mathcal{O}_{Cands}$ of candidate abnormal set of examples. Two are the desiderata that aCBOut must meet in order to deal with large set of examples, that are the set $\mathcal{O}_{Cands}$ should be of limited size, but anyway contain promising candidate abnormal sets, and it should be determined efficiently. Overall, the cost should be linearly related to the number $|\mathcal{E}|$ of examples and the exponential dependency from the parameter k_{max} should be broken.

Figure 9 reports the *a*CBOut-*candidate-selector* function. *a*CBOut selects candidate abnormal sets by exploiting ϱ-consistent hypotheses in order to fast detect candidate abnormal sets. Given a ϱ-consistent hypothesis $\mathcal{H}_\varrho^\mathcal{E}$ (with $\varrho > 0$), the candidate abnormal sets are selected among the sets of negative examples $covers_{\mathcal{E}^-}(c)$ covered by the clauses c of $\mathcal{H}_\varrho^\mathcal{E}$. Moreover, in order to rank these sets with respect to their significance as abnormal sets, a score is assigned to each of them. Specifically, the score associated with the set $covers_{\mathcal{E}^-}(c)$ is determined as follows:

$$\frac{\max_{c' \in \mathcal{H}_{\varrho-1}^\mathcal{E}} |covers_{\mathcal{E}^+}(c) \setminus covers_{\mathcal{E}^+}(c')| + 1}{\frac{1}{|\mathcal{H}_\varrho^\mathcal{E}|-1} \sum_{c'' \in \mathcal{H}_\varrho^\mathcal{E} \setminus \{c\}} |covers_{\mathcal{E}^+}(c'')|}.$$

Intuitively, this score measures to what extent the removal of the examples in $covers_{\mathcal{E}^-}(c)$ contributes to ameliorate the generalization of the examples in $covers_{\mathcal{E}^+}(c)$.

For now, take into account only the numerator of the score. If the examples in $covers_{\mathcal{E}^+}(c)$ are already covered all together by a clause c' of $\mathcal{H}_{\varrho-1}^\mathcal{E}$, then removing the examples in $covers_{\mathcal{E}^-}(c)$ does not help further in generalizing the concept underlying the examples in $covers_{\mathcal{E}^+}(c)$, and in this case the numerator evaluates to one (i.e., the minimum value). On the contrary, if no pair of examples from $covers_{\mathcal{E}^+}(c)$ is also included in a clause c' of $\mathcal{H}_{\varrho-1}^\mathcal{E}$, then removing the examples in $covers_{\mathcal{E}^-}(c)$ allows to induce a rule covering $|covers_{\mathcal{E}^+}(c)|$ previously (that is, according to the $(\varrho-1)$-consistent hypothesis) uncorrelated examples, and in this case the numerator evaluates to $|covers_{\mathcal{E}^+}(c)|$ (the maximum possible value for the considered examples).

As for the denominator, it represents the mean number of examples covered by the clauses belonging to the current hypothesis $\mathcal{H}_\varrho^\mathcal{E}$ after having excluded the clause c. Its role is to mitigate the bias towards (possibly larger) candidate sets associated with large values of the parameter ϱ.

The candidate abnormal sets are the sets scoring the m largest abnormality scores among those associated with clauses belonging to ϱ-consistent hypotheses with $\varrho \leq k_{max}$, where m and k_{max} are two user-provided parameters. Note that the same set could be associated with different clauses. In such a case, the best score associated with that set has to be considered.

Both positive and negative candidate abnormal sets are determined by considering alternatively the direct (for negative candidates) and the dual concept (for positive candidates). The candidate sets returned in the set $\mathcal{O}_{Cands}$ by the *a*CBOut-*candidate-selector* function are then checked for abnormality by computing their compliance both with $\mathcal{E} \cup \mathcal{B}$ and $\overline{\mathcal{E}} \cup \mathcal{B}$ and, finally, the minimal irregular, anomalous, and outlier sets in $\mathcal{O}_{Cands}$ form the output of the *a*CBOut algorithm.

Notice that, due to the way candidate abnormal sets are selected, *a*CBOut is primarily interested in anomalous and outlier sets. This should not be seen as a limitation, but rather as a peculiarity of the method, since anomalous and outliers sets are the subtler form of abnormality here considered and, moreover, since the algorithm can discover also irregulars.

Consider now the temporal cost of aCBOut. Computing the abnormality score costs $O(h|\mathcal{E}|)$, where h is the maximum number of clauses in an induced hypothesis and $O(|\mathcal{E}|)$ accounts for the cost of computing the difference between two sets. Thus, the cost of the aCBOut-*candidate-selector* function is $O(k_{max} \cdot (C_{ind}(\mathcal{B}, \mathcal{E}) + h^2|\mathcal{E}|))$ and, as a whole, the cost of the aCBOut algorithm is

$$O(k_{max} \cdot (C_{ind}(\mathcal{B}, \mathcal{E}) + h^2|\mathcal{E}|) + |\mathcal{O}_{Cands}| \cdot (C_{ind}(\mathcal{B}, \mathcal{E}) + h|\mathcal{E}|)),$$

with $|\mathcal{O}_{Cands}|$ not greater than $2m$. It can be concluded that the temporal cost of aCBOut makes it suitable for detecting abnormal sets in presence of large set of examples.

The only term showing a quadratic dependence is h, the maximum number of clauses of an hypothesis. However, it must be noted that in practice h is small. Moreover, there is a trade-off between the number of clauses h and the average number of examples covered per clause, so that if h is not very small, then it is the case that there is little overlap between clauses and computing the symmetric difference is easier. Thus, in practice the cost associated with score computation is far lower than the upper bound $h^2|\mathcal{E}|$.

Notice further that aCBOut can take fully advantage of parallelization, both during the first phase of candidate computation, when scores or score terms could be computed separately, and during the second phase of gain computation, when the compliance relationships could be checked apart. A similar strategy can be used also to parallelize CBOut and hCBOut, but for these algorithms some dependencies among candidate sets have to be taken into account that lower their degree of parallelization with respect to that of aCBOut.

Before concluding, we remark that both aCBOut and hCBOut are consistent and, since they explore only a portion of the search space, are necessarily incomplete. Clearly, the reduced cost of aCBOut has a counterpart. Since it does not consider the downward closure of the candidate abnormal sets (for otherwise the cost would turn to be exponential in the parameter k_{max}) the returned sets could not be minimal with respect the containment between abnormal sets. However, the strategy employed to select candidates is supposed to provide reasonably fast good abnormals even in presence of approximate computations.

6 Related Work

Many ILP systems include mechanisms for handling *imperfect data* in order to make ILP applicable to real-life problems. In particular, according to the taxonomy mentioned in (Lavrač et al, 1996), the following forms of imperfect data can be encountered: *random errors*, or *noise*, that is either noise in the examples (caused by erroneous argument values and/or erroneous classification of facts as true or false) or noise in the background knowledge; *incompleteness*,

that is to say too sparse examples for which it is difficult to reliably detect regularities; *inappropriateness*, that is to say imperfect background knowledge, due either to predicates that are not relevant for the learning task or to predicates insufficient for learning; *missing values*, that are missing argument values in examples. ILP learning systems usually have a single mechanism, called *noise-handling mechanism*, for dealing with the first three kinds of imperfect data (noisy, incomplete and inexact data), which prevents the induced hypothesis from overfitting the data set, while missing values are usually handled by a separate mechanism. The noise-handling mechanisms are of two types: using appropriate search heuristics and stopping criteria during the hypothesis construction; generating the target predicate definition as the data were completely correct by employing the standard consistency and completeness stopping criteria and, then, post-processing the induced hypothesis.

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. We recall that the problem definition given in Section 2.2 requires that the induced hypothesis is *complete* and *consistent*. We point out that noise possibly present in the examples does not affect the task of detecting abnormals, since the noise is retained both in presence and in absence of the set to be tested for abnormality. Moreover, the notions of starting and ending theories, other than playing the important role of explanation, may greatly mitigate the influence on the portions of hypotheses compared when testing the compliance relationship of most of the noise possibly present. Furthermore, the task considered here is a knowledge discovery one. The knowledge mined by the method is far richer than noise. In particular, as already discussed in Section 3, anomalous sets are very different from noise, since they 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. As far as irregular and outlier sets is concerned, it can be said that they have some analogies with noise in examples, since the description of the concept would be significantly more concise if each example in the set were not observed. Nonetheless, the distinction between irregular sets and outlier sets allows us to provide a finer characterization of the abnormality at hand. In particular, as previously noted, while we can imagine that irregular sets are "far away" the majority of the positive examples, but anyway "far" from the negative ones, differently we can imagine that outlier sets lie very close or within the "shape" of the dual concept.

In (Angiulli et al, 2007, 2008) a notion of outlier in the context of some knowledge-based systems is defined. In particular, the formal frameworks considered are those of *Default Logics* and *Extended Disjunctive Logic Programming* under answer set semantics (Angiulli et al, 2008) and *Logic Programming* under stable model semantics (Angiulli et al, 2007). Loosely speaking, given a logic program P and a set of facts F, a subset O of F is said to be an *outlier*, if there exists a nonempty set W of F, called *witness*, such that the two following conditions hold: (i) $P \cup (F \setminus W) \models \neg W$, and (ii) $P \cup (F \setminus (W \cup O)) \not\models \neg W$,

where $\neg W$ denotes the conjunction of the negation of the atoms occurring in W, that is $\neg W = \neg w_1 \wedge \ldots \wedge \neg w_n$ with $W = \{w_1, \ldots, w_n\}$. Intuitively, the first condition states that the facts in W encode an unexpected property, since if these observations were dropped from the set of facts at hand then the exact opposite would have been concluded, while the second condition states that the unexpected property W is indeed a property related to the facts in O, which hence encode knowledge exhibiting some anomalous behavior.

There are major differences between the approach presented here and those above recalled. First, our approach is based on *induction*, while the above mentioned approaches 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 (Angiulli et al, 2007, 2008) it is perceived by means of the satisfaction of certain conditions involving the entailment operator. Moreover, importantly, the definitions proposed in (Angiulli et al, 2007, 2008) 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 (Angiulli et al, 2007) *there are no outliers in a logic program* (cf. Theorem 3.2 of (Angiulli et al, 2007)). In fact, if $P \cup (F \setminus (W \cup O)) \not\models \neg W$, since the program is positive it is the case that $P \cup (F \setminus (W \cup O)) \models W$ and, hence, by monotonicity, also that $P \cup (F \setminus W) \models W$. This makes the two approaches for defining outliers incomparable from a practical point of view.

Somehow related to the research conducted here are the anomaly detection techniques exploiting *approximate dependency-based methods*. These techniques are based on the inference of functional dependencies from data. Mannila and Räihä (1987) deal with the problem of inferring functional dependencies from data. A functional dependency (FD) is a rule of the form $A \rightarrow b$ where A is a set of attributes and b is an attribute stating that if the dataset objects assume the same value on the set of attribute A they must assume the same value also on the attribute b. Initially, these kind of rules have been introduced and extensively used as user-defined constraints for defining a database schema. Afterwards, FDs have been seen as pieces of knowledge that could be mined from data: inferring from a dataset an unexpected FD could highlight an interesting property. In this latter sense, some efforts have been made to interpret FD inference from a data mining point of view (Novelli and Cicchetti, 2001). It is quite clear that, in real dataset, often an FD does not strictly hold, since the property the FD represents could be valid from *most* the dataset objects but not for *all* the dataset objects. As a consequence, especially for data mining purposes, it is needed to infer *approximate functional dependencies* (A-FD), namely FD holding from *most* the dataset objects, both in relational database (Kivinen and Mannila, 1995) and in XML database (Fassetti and Fazzinga, 2007). Quite naturally, the objects not satisfying the inferred A-FDs could be seen as a particular kind of "anomaly",

as highlighted by (Bruno et al, 2007) in the XML context. Even if the idea pursued in these approaches concerns the inference of a kind of rule and the detection of the anomalies is made on the basis of the inferred rule, there are substantial differences with our approach. First of all, the A-FD based techniques are able to identify only objects that are abnormal since they do not fit the positive class (namely, irregular objects), then they are not able to mine anomalous or outlier objects. Moreover, these techniques cannot take advantages from the presence of a background knowledge which could carry relevant information for the analysis. Finally, they do not supply the abnormal objects with an explanation of the characterizing abnormality.

Outlier detection in data mining considers the following task: "Given a set of data points or objects, find the objects that are considerably dissimilar, exceptional or inconsistent with respect to the remaining data". Early methods for outlier identification have been developed in the field of statistics. Different outlier detection approaches have been proposed in the literature, as *distance-based* (Knorr and Ng, 1998; Angiulli and Pizzuti, 2002; Ramaswamy et al, 2000; Angiulli et al, 2006; Angiulli and Fassetti, 2009a), *density-based* (Breunig et al, 2000; Papadimitriou et al, 2003), *frequent pattern*-based (He et al, 2005), *projection*-based (Aggarwal and Yu, 2001), *angle-based* (Kriegel et al, 2008), *isolation-based* (Liu et al, 2012), and others (Chandola et al, 2009). Among these approaches, distance-based outlier detection has been introduced by Knorr and Ng (1998) to overcome the limitations of statistical methods: an object O is a distance-based outlier in a data set with respect to parameters k and R if at least k objects in the data set lie within distance R from O. This definition generalizes the definition of outlier in statistics. Moreover, it is suitable in situations when the data set does not fit any standard distribution. The assumption of distance-based methods is that it is possible to compute for each pair of objects their distance. First-order distance measures were proposed and used in various distance-based multi-relational algorithms (Kirsten et al, 2001), such as RIBL2, an instance-based learner applying the k-nearest neighbor classifier, RDBC, a hierarchical agglomerative clustering method, and FORC, a k-means clustering algorithm.

Clearly, there are major differences between the approach here proposed and these data analysis tools. Almost all of them are not able to deal with labeled examples. Moreover, outliers returned by these methods, and also by almost all of the unsupervised outlier detection approaches, are of different nature with respect to those returned by the approach here introduced. Intuitively, these outliers are likely to correspond to irregular instances as defined here, since they are individuals whose attribute-value pattern is shared less.

7 Experimental results

In this section we present experiments conducted by using the outlier detection approach here introduced.

We implemented the CBOut algorithm and its variants in Yap Prolog on top of the P-Progol system[1] which is based on the PROGOL algorithm (Muggleton, 1995). Progol combines *inverse entailment* with *general-to-specific search* through a refinement graph. Inverse entailment is used with mode declarations to derive the most-specific clause within the mode language which entails a given example. This clause is used to guide a refinement-graph search. Progol's search is efficient and has a provable guarantee of returning a solution having the maximum compression in the search-space. To do so it performs an admissible A*-like search, guided by compression, over clauses which subsume the most specific clause.

Experiments are organized as follows. Section 7.1 discusses the kind of knowledge mined by CBOut in different contexts. Section 7.2 shows how to improve efficiency while maintaining good accuracy in domains with a few hundreds of examples and possibly large background theories, by exploiting the *h*CBOut algorithm. Finally, Section 7.3 illustrates the application of the here introduced approach to large set of examples, by means of *a*CBOut.

7.1 Knowledge mined

This section discusses the kind of knowledge discovered by CBOut in different contexts. Specifically, Section 7.1.1 considers a zoo data set containing instances associated with animals and their properties, subsequent Section 7.1.2 takes into account a student loan relational domain, Section 7.1.3 explores a mutagenesis data set concerning the prediciton of carcinogenesis and, finally, Section 7.1.4 compares the kind of knowledge singled out by CBOut with distance-based outliers.

7.1.1 Zoo data set

In this experiment we considered the Zoo data set from the UCI Machine Learning Repository[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* (a value in the set {0, 2, 4, 5, 6, 8}), and the following boolean attributes: *hair*, *feathers*, *eggs*, *milk*, *airbone*, *aquatic*, *predator*, *toothed*, *backbone*, *breathes*, *venomous*, *fins*, *tail*, *domestic*, *catsize*.

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.

[1] `www.comlab.ox.ac.uk/oucl/research/areas/machlearn/PProgol/pprogol.pl`.

[2] `http://archive.ics.uci.edu/ml/datasets/Zoo`.

We executed the algorithm with $\alpha = 0.05$ and $k_{max} = 1$. 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:

- $\mathcal{O}_1 = \{class(amphibian, newt)\}$ as positive outlier,
- $\mathcal{O}_2 = \{class(insect, ladybird)\}$ as positive anomalous, and
- $\mathcal{O}_3 = \{class(mammal, platypus)\}$ as positive anomalous.

Next we comment on some of the knowledge discovered by the method.

The positive outlier set $\mathcal{O}_1 = \{class(amphibian, newt)\}$ is a fact in the direct theory, while it has as dual explanation the dual starting theory $\overrightarrow{\mathcal{H}}^{\overline{\varepsilon}}_{\mathcal{B}}(\overline{\mathcal{O}}_1)$:

$$not\_class(amphibian, X) \leftarrow catsize(X) \ \ \langle 44 \rangle$$
$$not\_class(amphibian, X) \leftarrow legs(2, X) \ \ \langle 27 \rangle$$
$$not\_class(amphibian, tuatara) \ \ \langle 1 \rangle$$
$$not\_class(amphibian, scorpion) \ \ \langle 1 \rangle,$$

and the dual ending theory $\overleftarrow{\mathcal{H}}^{\overline{\varepsilon}}_{\mathcal{B}}(\overline{\mathcal{O}}_1)$:

$$not\_class(amphibian, X) \leftarrow tail(X) \ \ \langle 74 \rangle,$$

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.

The positive anomalous set $\mathcal{O}_2 = \{class(insect, ladybird)\}$ has as dual explanation the dual starting theory $\overrightarrow{\mathcal{H}}^{\overline{\varepsilon}}_{\mathcal{B}}(\overline{\mathcal{O}}_2)$:

$$not\_class(insect, X) \leftarrow aquatic(X) \ \ \langle 35 \rangle$$
$$not\_class(insect, scorpion) \ \ \langle 1 \rangle,$$

and the dual ending theory $\overleftarrow{\mathcal{H}}^{\overline{\varepsilon}}_{\mathcal{B}}(\overline{\mathcal{O}}_2)$:

$$not\_class(insect, X) \leftarrow predator(X) \ \ \langle 54 \rangle,$$

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.

The positive anomalous set $\mathcal{O}_3 = \{class(mammal, platypus))\}$ has as dual explanation the dual starting theory $\overrightarrow{\mathcal{H}}^{\overline{\mathcal{E}}}_{\mathcal{B}}(\overline{\mathcal{O}}_3)$:

$$not\_class(mammal, X) \leftarrow legs(6, X) \ \ \langle 10 \rangle$$
$$not\_class(mammal, X) \leftarrow feathers(X) \ \ \langle 20 \rangle$$
$$not\_class(mammal, X) \leftarrow eggs(X), toothed(X) \ \ \langle 19 \rangle$$
$$not\_class(mammal, X) \leftarrow eggs(X), legs(0, X) \ \ \langle 19 \rangle$$
$$not\_class(mammal, starfish) \ \ \langle 1 \rangle$$
$$not\_class(mammal, tortoise) \ \ \langle 1 \rangle$$
$$not\_class(mammal, crab) \ \ \langle 1 \rangle$$
$$not\_class(mammal, octopus) \ \ \langle 1 \rangle,$$

and the dual ending theory $\overleftarrow{\mathcal{H}}^{\overline{\mathcal{E}}}_{\mathcal{B}}(\overline{\mathcal{O}}_3)$:

$$not\_class(mammal, X) \leftarrow eggs(X) \ \ \langle 57 \rangle,$$

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.

7.1.2 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 $k_{max} = 1$, with 78 positive examples and 34 negative examples, consisting of the students whose identifier number starts with 1.

Besides the facts in the induced hypothesis and the induced dual hypothesis, the CBOut algorithm reported the following abnormal sets:

- $\mathcal{O}_1 = \{no\_payment\_due(student149)\}$ as negative outlier,
- $\mathcal{O}_2 = \{no\_payment\_due(student116)\}$ as negative outlier, and
- $\mathcal{O}_3 = \{no\_payment\_due(student102)\}$ as positive anomalous.

Next we briefly comment on some knowledge discovered by the method. In the following we will denote by *payment_due* the predicate *not_no_payment_due* associated with the dual concept.

The negative outlier set $\mathcal{O}_1 = \{no\_payment\_due(student149)\}$ has, as direct explanation, the direct starting theory $\overrightarrow{\mathcal{H}}^{\mathcal{E}}_{\mathcal{B}}(\mathcal{O}_1)$:

$$no\_payment\_due(X) \leftarrow enrolled(X, Y, 10) \ \ \langle 12 \rangle,$$

[3] `http://archive.ics.uci.edu/ml`.

and the direct ending theory $\overleftarrow{\mathcal{H}}^{\mathcal{E}}_{\mathcal{B}}(\mathcal{O}_1)$:

$$no\_payment\_due(X) \leftarrow male(X), enrolled(X, Y, 10) \ \ \langle 6 \rangle$$
$$no\_payment\_due(student165) \ \ \langle 1 \rangle$$
$$no\_payment\_due(student112) \ \ \langle 1 \rangle$$
$$no\_payment\_due(student196) \ \ \langle 1 \rangle,$$

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

The positive anomalous set $\mathcal{O}_3 = \{no\_payment\_due(student102)\}$ has, as dual explanation, the dual starting theory $\overrightarrow{\mathcal{H}}^{\mathcal{E}}_{\mathcal{B}}(\overline{\mathcal{O}_3})$:

$$payment\_due(X) \leftarrow male(X), enrolled(X, Y, 3) \ \ \langle 6 \rangle,$$

and the dual ending theory $\overleftarrow{\mathcal{H}}^{\mathcal{E}}_{\mathcal{B}}(\overline{\mathcal{O}_3})$:

$$payment\_due(X) \leftarrow longest\_absence\_from\_school(X, 7), enrolled(X, Y, 3) \ \ \langle 2 \rangle$$
$$payment\_due(103) \ \ \langle 1 \rangle$$
$$payment\_due(110) \ \ \langle 1 \rangle$$
$$payment\_due(111) \ \ \langle 1 \rangle$$
$$payment\_due(180) \ \ \langle 1 \rangle,$$

with gain 0.06. The *student*102 is the only male enrolled in three units which is not required to repay a student loan.

7.1.3 Mutagenesis data set

Mutagenesis is relevant to the understanding and the prediction of carcinogenesis. The problem here considered consists in predicting the mutagenicity of a set of aromatic and heteroaromatic nitro compounds by using only the atomic bond structure of the compounds. The data is based on the results in (Debnath et al, 1991) and comes from ILP experiments conducted with Progol and described in (Srinivasan et al, 1996).[4] Each compound is represented by a sets of facts of the form:

- $bond(compound, atom1, atom2, bondtype)$: stating that *compound* has a bond of *bondtype* between the atoms $atom1$ and $atom2$; and
- $atm(compound, atom, element, atomtype, charge)$: stating that in *compound* *atom* has *element* of *atomtype*.

The background knowledge consists in 12,203 facts on atomic structure and bonding plus some rules that define generic chemistry knowledge and concepts, as ring structures.

[4] Data are available at `http://www.comlab.ox.ac.uk/activities/machinelearning/mutagenesis.html`.

Of the 230 compounds, 138 have positive levels of log mutagenicity, these are labeled *active* and constitute the positive examples. The remaining 92 compounds are labeled *inactive* and constitute the negative examples. Moreover, the original (Debnath et al, 1991) paper recognized two subsets of data: 188 compounds that could be fitted using linear regression (125 active and 63 inactive), and 42 compounds that could not (13 active and 29 inactive). The model makes use of some independent variables, among which there are the *logp*, that is the hydrophobicity of the compound, and the *lumo*, that is the energy level of the lowest unoccupied molecular orbital.

We executed the algorithm CBOut on the set of examples corresponding to the compounds that fit linear regression in order to isolate abnormalities in the supposedly regular data. The parameters used are $\alpha = 0.05$ and $k_{max} = 1$. Besides the facts in the induced direct and dual hypotheses returned as irregular sets, the algorithm reported the following abnormal sets: $\mathcal{O}_1 = \{active(d65)\}$ as negative outlier, $\mathcal{O}_2 = \{active(d178)\}$ as positive anomalous, $\mathcal{O}_3 = \{active(d70)\}$ and $\mathcal{O}_4 = \{active(d188)\}$ as negative anomalous.

Next we comment on the knowledge associated with some of the discovered abnormal sets.

The negative outlier set $\mathcal{O}_1 = \{active(d65)\}$ is a fact in the dual theory while it has, as direct explanation, the direct starting theory $\overrightarrow{\mathcal{H}}^{\mathcal{E}}_{\mathcal{B}}(\mathcal{O}_1)$:

$active(X) \leftarrow atm(X, Y, o, 40, -0.389), bond(X, Z, Y, 2), bond(X, W, Z, 1)$ $\langle 11 \rangle$
$active(d105)$ $\langle 1 \rangle$

and the direct ending theory $\overleftarrow{\mathcal{H}}^{\mathcal{E}}_{\mathcal{B}}(\mathcal{O}_1)$:

$$active(X) \leftarrow atm(X, Y, c, 27, Z), lumo(X, W), W \leq -1.749 \quad \langle 42 \rangle$$

with gain 0.309. The knowledge discovered states that having the compound $d65$ as a negative example prevents to induce the following knowledge generalizing a large number of positive examples (37% of the positive examples): in the representation employed, a carbon atom of type 27 with *lumo* less than -1.749 is active.

The negative anomalous set $\mathcal{O}_4 = \{active(d188)\}$ has, as direct explanation, the direct starting theory $\overrightarrow{\mathcal{H}}^{\mathcal{E}}_{\mathcal{B}}(\mathcal{O}_4)$:

$active(X) \leftarrow atm(X, Y, c, 22, -0.122), logp(X, Z), Z \geq 2.74$ $\langle 14 \rangle$
$active(X) \leftarrow atm(X, Y, c, 22, -0.114), bond(X, Z, Y, 1)$ $\langle 7 \rangle$
$active(X) \leftarrow atm(X, Y, c, 10, Z), atm(X, W, c, 29, Z), ring\_size\_5(X, V)$ $\langle 15 \rangle$
$active(X) \leftarrow atm(X, Y, o, 40, -0.389), bond(X, Z, Y, 2), bond(X, W, Z, 1)$ $\langle 11 \rangle$
$active(d79)$ $\langle 1 \rangle$
$active(d18)$ $\langle 1 \rangle$
$active(d87)$ $\langle 1 \rangle$

and the direct ending theory $\overleftarrow{\mathcal{H}}^{\mathcal{E}}_{\mathcal{B}}(\mathcal{O}_4)$:

$$active(X) \leftarrow atm(X,Y,c,29,Z), ring\_size\_5(X,W) \ \ \langle 36\rangle$$
$$active(X) \leftarrow atm(X,Y,n,38,Z), Z \leq 0.794 \ \ \langle 10\rangle$$

with gain 0.120. Also in this case, having the compound $d188$ as a negative example prevents to induce the piece of knowledge associated with the ending theory which sensibly increases generalization.

The positive anomalous set $\mathcal{O}_2 = \{active(d178)\}$ has, as dual explanation, the dual starting theory $\overrightarrow{\mathcal{H}}^{\overline{\mathcal{E}}}_{\mathcal{B}}(\overline{\mathcal{O}_2})$:

$$not\_active(X) \leftarrow atm(X,Y,h,3,0.137), lumo(X,Z), Z \geq -1.266 \ \ \langle 7\rangle$$
$$not\_active(X) \leftarrow atm(X,Y,n,34,Z), atm(X,Y,c,21,W), W \geq -0.107 \ \ \langle 7\rangle$$
$$not\_active(X) \leftarrow atm(X,Y,o,50,Z), lumo(X,W), W \leq -1.474 \ \ \langle 6\rangle$$
$$not\_active(X) \leftarrow atm(X,Y,c,22,Z), Z \leq -0.174 \ \ \langle 2\rangle$$
$$not\_active(d168) \ \ \langle 1\rangle$$
$$not\_active(d88) \ \ \langle 1\rangle$$

and the dual ending theory $\overleftarrow{\mathcal{H}}^{\overline{\mathcal{E}}}_{\mathcal{B}}(\overline{\mathcal{O}_2})$:

$$not\_active(X) \leftarrow lumo(X,Y), Y \geq -1.24, methyl(X,Z) \ \ \langle 12\rangle$$
$$not\_active(X) \leftarrow logp(X,Y), Y \leq 1.87, methyl(X,Z) \ \ \langle 10\rangle$$

with gain 0.128, similar to that of $\mathcal{O}_4$.

Moreover, we re-executed the algorithm CBOut on the whole set of examples with parameters $\alpha = 0.05$ and $k_{max} = 2$ in order to isolate abnormal sets composed of at most two elements. As an example, CBOut returned the the negative outlier set $\mathcal{O}' = \{active(d192), active(d193)\}$. This set has as direct explanation, the starting theory $\overrightarrow{\mathcal{H}}^{\mathcal{E}}_{\mathcal{B}}(\mathcal{O}')$:

$$active(X) \leftarrow atm(X,Y,o,40,-0.384), logp(X,Z), Z \geq 2.06 \ \ \langle 11\rangle$$

and the ending theory $\overleftarrow{\mathcal{H}}^{\mathcal{E}}_{\mathcal{B}}(\mathcal{O}')$:

$$active(d122) \ \ \langle 1\rangle$$

that means that the negative examples in $\mathcal{O}'$ prevent to induce the clause in the starting theory. We note that if just one of the two examples in $\mathcal{O}'$ is removed from the set of negative examples the clause in the starting theory cannot be induced. In the dual hypothesis, the two examples in $\mathcal{O}'$ are covered by a clause covering just them, and, as a result, $\overline{\mathcal{O}'}$ does not comply with $\overline{\mathcal{E}} \cup \mathcal{B}$. Interestingly, this pair of examples belongs to the set of examples that does not fit the linear regression model.

7.1.4 Comparison with Distance-Based Outliers

Here, the knowledge mined by CBOut is compared with distance distance-based outliers. The distance-based outliers in the original *Zoo* data set are detected by using as outlier score the sum of the distances to the k-nearest neighbors (Angiulli and Pizzuti, 2005) and the Hamming function as distance measure. 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:

Outlier (Score)	Nearest Neighbors
1. *scorpion* (25)	$worm, slug, pitviper, clam, crab$
2. *seasnake* (19)	$pitviper, stingray, chub, herring, bass$
3. *tortoise* (18)	$tuatara, ostrich, rhea, slowworm, wren$
4. *toad* (17)	$frog, newt, tuatara, worm, crab$
5. *pitviper* (15)	$slowworm, tuatara, seasnake, newt, kiwi$

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 by irregular sets. The example $class(invertebrate, scorpion)$ 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.

7.2 Improving efficiency

In this section it is shown how to improve efficiency of CBOut by taking advantage of the hCBOut algorithm in domains where a few hundreds of examples are available possibly together with a complex background theory.

Figure 10 shows the execution time[5] (in seconds) of the second phase of the hCBOut algorithm on the *Zoo* (Figure 10a), *Student loan* (Figure 10b), *Mutagenesis regression friendly* (consisting of the compounds that can be fitted by linear regression (Figure 10c), and *Mutagenesis* data sets (Figure 10d).

Various values for the parameter m, that are $m = 5$ (the solid lower curve, with triangles), $m = 10$ (the solid middle curve, with squares), and $m = 20$ (the solid upper curve, with circles), and for the parameter k_{max} varying up to either 3 or 5, have been considered.

[5] We employed Intel Xeon E5620 2.40GHz based computer with 4GB of main memory and the Linux operating system.

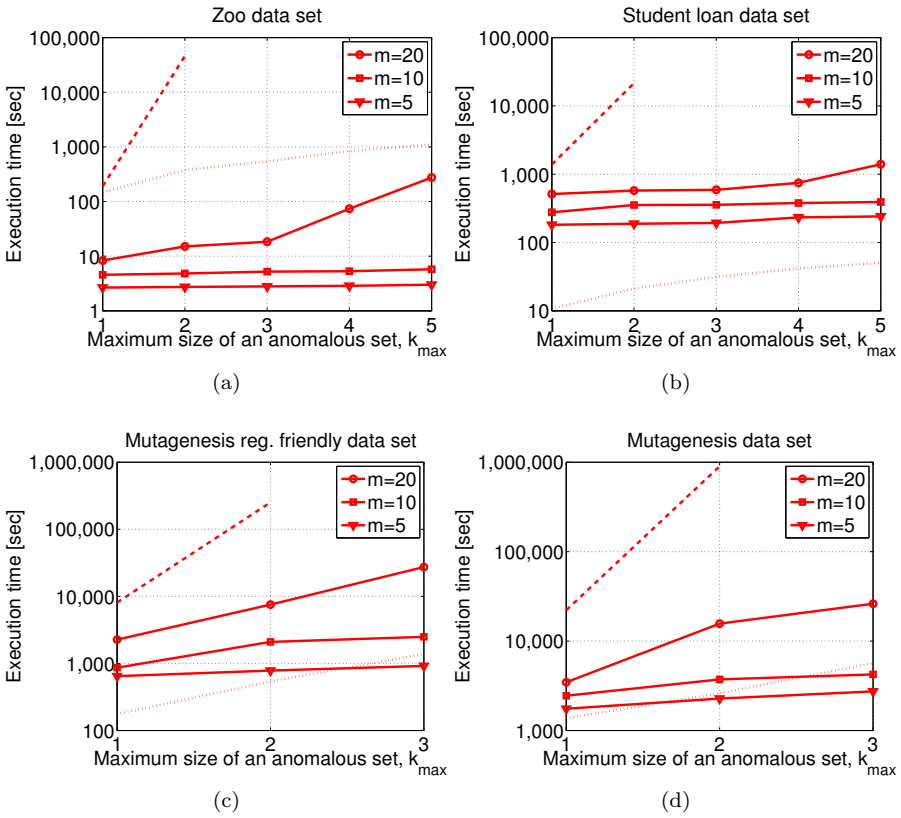


Fig. 10: Scalability of the hCBOut algorithm.

The following table summarizes the size of the sets of examples and of the background theories associated with the four data sets.

Data set	$\lvert\mathcal{E}^+\rvert$	$\lvert\mathcal{E}^-\rvert$	$\mathcal{B}$	
			facts	rules
Zoo	100	600	1,713	0
Student loan	78	34	521	16
Mutagenesis reg. friendly	125	63	15,040	41
Mutagenesis	138	92	15,040	41

¿From Figure 10, it is clear that the execution time increases with the parameter m. As far as the curves for $m = 5$ and $m = 10$ are concerned, in almost all the experiments the execution time appears to be little influenced by the parameter k_{max}. As for the curve with $m = 20$, the dependence of the execution time from the parameter k_{max} appears to be more evident, since in this case the number of candidate sets to explore sensibly increases.

Data set	m	*Candidates*		*Recall*
		Abnormal	*Normal*	
Zoo	5	100%	0%	35.1%
	10	100%	0%	62.2%
	20	85.7%	14.3%	97.3%
Student loan	5	100%	0%	32.6%
	10	100%	0%	51.2%
	20	90.0%	10.0%	86.1%
Mutagenesis reg. friendly	5	93.3%	6.7%	42.4%
	10	66.7%	33.3%	48.5%
	20	58.5%	41.5%	72.7%
Mutagenesis	5	93.8%	6.2%	30.0%
	10	96.0%	4.0%	48.0%
	20	81.8%	18.2%	72.0%

Table 2: *Recall* of the hCBOut algorithm.

The dotted curves represent the execution time of the first phase of the hCBOut algorithm, that is the candidate selection phase. In general, the cost of the first phase is either comparable or noticeably smaller than that of the second phase. On the *Zoo* data set the first phase is more costly than the second one, since the cost of inducing an hypothesis on this data set is small due to the simplicity of the associated background theory. On the other data sets, for $m = 20$ the cost of the first phase is negligible with respect to the cost of the second one.

In any case, it must be noticed that the candidate selection phase guarantees to the hCBOut algorithm vast time savings with respect to the CBOut one. Indeed, consider the dashed curve, corresponding to the execution time of CBOut (only the values $k_{max} \in \{1, 2\}$ have been considered, since for greater values this algorithm requires too much time). Notice that for $k_{max} = 2$ the algorithm CBOut requires about half a day on *Zoo*, six hours on *Student loan*, three days on *Mutagenesis reg. friendly*, and ten days on *Mutagenesis*, while hCBOut terminates after about fifteen minutes on the first two data sets, two hours and half on the third one, and five hours on the last data set.

Table 2 reports the *Recall* and the fractions of normal and abnormal candidates returned by the hCBOut algorithm measured for $k_{max} = 1$. Let $\mathcal{A}$ denote the set of examples belonging to at least an abnormal set. The fraction of candidate examples in $\mathcal{E}_{Cand}$ which are actually part of an abnormal set is

$$\frac{|\mathcal{E}_{Cand} \cap \mathcal{A}|}{|\mathcal{E}_{Cand}|},$$

while the fraction of those which are not part of an abnormal set is $\frac{|\mathcal{E}_{Cand} \setminus \mathcal{A}|}{|\mathcal{E}_{Cand}|}$. The *Recall* is the fraction of examples that are actually part of an abnormal set belonging to the set $|\mathcal{E}_{Cand}|$, that is

$$Recall = \frac{|\mathcal{E}_{Cand} \cap \mathcal{A}|}{|\mathcal{A}|}.$$

From Table 2 it is clear that the recall (last column) increases with the parameter m. We notice that the size of the set $\mathcal{E}_{Cands}$ is directly proportional to the value of the parameter m. Thus, the larger the parameter m, the larger the size of $\mathcal{E}_{Cands}$, and the greater the chance for the set $\mathcal{E}_{Cand}$ to accommodate abnormal examples. The fraction of examples in $\mathcal{E}_{Cands}$ that are actually abnormal ones (the third column: abnormal candidates) shows that the quality of the candidates selected by hCBOut is very good. For small values of m ($m \in \{5, 10\}$) almost all the candidates are abnormal. For $m = 20$, there are candidate examples which are not abnormal, but the recall greatly increases (up to the 97.3% for *Zoo*, the 86.1% for *Student loan*, and the 72.0% for *Mutagenesis*).

Compare the two versions of the *Mutagenesis* data set; although the recall is similar, it appears that a greater number of normal examples are selected as candidates on the regression friendly data set. Since the *Mutagenesis reg. friendly* data set has been obtained by the whole *Mutagenesis* one by removing "outliers", it appears that the hCBOut is effective in isolating examples exhibiting a clear abnormality.

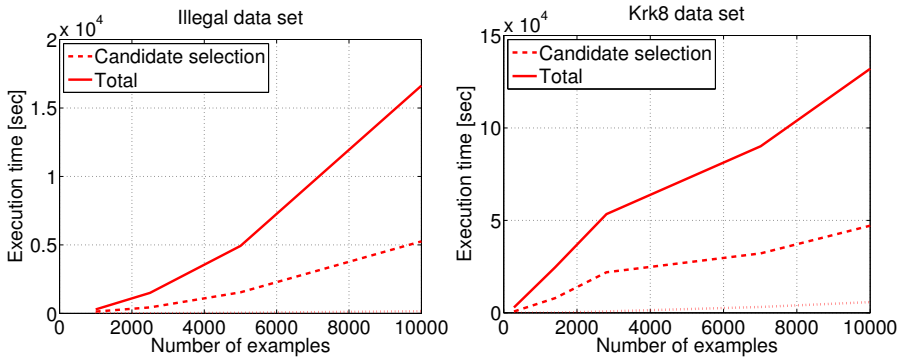
7.3 Experiments on large sets of examples

In this section, experiments on data sets containing a large set of examples are described. The huge dimension of the domains here considered can be efficiently faced only with the aCBOut algorithm. Hence, experiments presented in the following make use of this variant of CBOut. Experiments are organized as follows. Section 7.3.1 describes the domains employed. Section 7.3.2 studies scalability of the approach. Section 7.3.3 discusses on the accuracy and the knowledge mined. Finally, Section 7.3.4 considers the task of comparing accuracy performance with and without abnormal instances detected on a family of synthetically generated noisy data sets.

7.3.1 Description of the domains

The domains explored in the experiments presented in this section are *Illegal* and *Krk*, both concerning the problem of learning chess endgames.

The *Illegal* domain (Muggleton et al, 1989) has become a widely accepted test-bed for ILP systems. It regards the King-Rook-King (KRK) chess endgame, corresponding to the situation in which only the following three pieces are left on the chess board: White King (WK), White Rook (WK), and Black King (BK). In particular, *Illegal* models the problem of learning rules for recognizing illegal positions when it is white's turn to move (WTM). *Illegal* consists of 3,241 positive examples and 6,760 negative examples. Examples are represented by the predicate *illegal* having six arguments standing for the file (column) and rank (row) coordinates for the WK, WR and BK. The background knowledge provided is the ordering of rows or columns on a chess board. The binary predicate *lt* tabulates pairs row or column values where one is less than



(a) (b)

Fig. 11: Scalability of the aCBOut algorithm.

the other. The binary predicate *adj* tabulates row or column values that are adjacent.

The *Krk* domain (Bain and Srinivasan, 1995) deals with the problem of predicting the optimal depth of win (that is, the number of moves to checkmate) in the KRK endgame. It consists of the exhaustive database for the KRK domain, where each example has associated with it optimal depth of win information. Here this is represented by the predicate *krk*. The seven arguments for this predicate stand for depth of win, and the file (column) and rank (row) coordinates for the WK, WR and BK. The total number of examples is 20,056. These examples can be used to learn, e.g., the sub-concept 'black-to-move KRK position won optimally for white (with)in N moves". The background knowledge consists of the ternary predicate *diff*, representing the "symmetric difference" between files and ranks, and of the binary predicate *lt* ("strictly less than") for all unordered pairs of the file values $A, \ldots, H$, and all unordered pairs of the rank values $1, \ldots, 8$.

If not otherwise stated, in the following $k_{max} = 5$ and $m = 10$ are employed as parameters of aCBOut.

7.3.2 Scalability

In this experiment, we considered the *Illegal* and *Krk8* data sets.

The *Krk8* data set has as been obtained from *Krk* by using as set of positive examples the KRK positions for which the White wins in at most height moves, and as set of negative examples all the remaining KRK positions (that is, either the White has a win between nine and sixteen moves or there is a draw). This data set consists of 3,809 positive examples and 24,247 negative ones.

In order to study the scalability, from each data set we generated a family of increasing size data sets preserving the class distribution. Figure 11 shows the scalability of the aCBOut algorithm. In particular, Figures 11a and 11b

show the execution time as a function of the number of examples taken into account. The solid curve represents the total execution time, the dashed curve the execution time of the candidate selection phase, while the dotted curve is the time spent by the method to compute the scores used for ranking candidate outlier sets.

Since the difference between the dashed and the dotted curve represents the time needed to compute the ϱ-consistent hypotheses needed to single out candidates, it can be seen that the computational effort is mostly associated with the task of computing hypotheses, which is heavy on these domains. Particularly, the KRK data set is much more difficult, since the distribution of its examples is very unbalanced and the induction algorithm is particularly slow during the induction of the hypothesis associated with the dual concept. ¿From the curves it can be observed that the algorithm processes in reasonable amount of time some thousands of examples. Its execution time is negatively influenced by unbalanced set of examples, due to the necessity of computing both direct and dual hypotheses. For data sets composed of tens of thousands examples, exploiting parallelization is definitively needed.

7.3.3 Knowledge mined

In this section we discuss on the kind of knowledge mined by the aCBOut algorithm.

Specifically, here we considered the *Krk_sl* data set (for KRK short vs long endgames), which has as set of positive examples the KRK positions for which the white wins in at most five moves (capturing the concept "the endgame terminates with a win in a few number of moves") and as set of negative examples the KRK positions for which the white wins in more than fifteen moves (capturing the concept "the endgame terminates with a win after a large number of moves"). The data set is composed of 1,101 positive examples and 2,556 negative ones, for a total of 3,675 examples.

Figure 12 concerns the quality of the candidates selected by aCBOut. On the abscissa it is reported a gain value α, while on the ordinate the percentage of candidate sets that do not α-comply with the direct or dual hypothesis (that is, that are classified as abnormal sets for that value of α). More than the half of the candidate examples form an abnormal set at the level $\alpha = 0.05$ (corresponding to a coverage increase of the 5% of the example set size). For $\alpha = 0.01$ (corresponding to a coverage increase of the 1%) all the candidate sets are to be considered abnormal. There are no candidates with gain below this threshold. From the analysis, it appears that the set of candidates selected by aCBOut is of remarkable quality and, thus, the method can be effectively exploited to mine abnormal sets in large domains.

Next we discuss some notable abnormal sets returned by aCBOut.

The positive example set $\mathcal{O}_1 = \{krk\_sl(c,1,d,4,a,3),\ krk\_sl(c,1,h,4,a,3),\ krk\_sl(c,1,e,4,a,3),\ krk\_sl(c,1,g,4,a,3),\ krk\_sl(c,1,f,4,a,3)\}$ is an outlier.

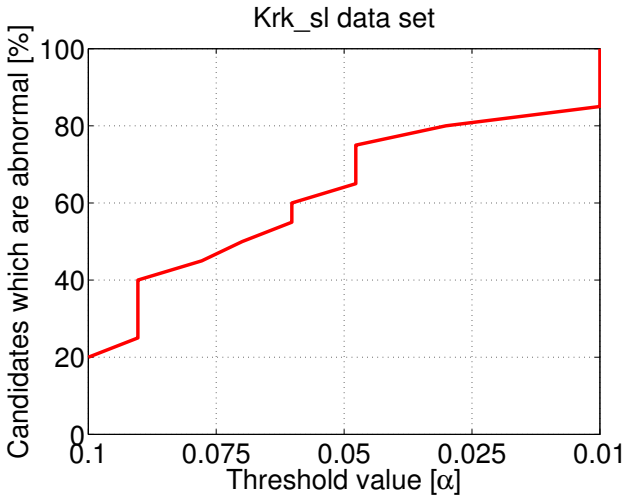


Fig. 12: Accuracy of aCBOut on the *Krk_sl* data set.

The associated dual starting theory is

$$not\_krk\_sl(A,B,C,D,E,F) \leftarrow edge(B), diff(B,F,G), diff(A,E,G) \quad \langle 325 \rangle,$$

a rule covering 325 negative examples. The corresponding dual ending theory is quite involved, consisting of 155 facts and three rules with relatively low coverage. The gain in this case is sensibly large, amounting to 0.127. Moreover, the examples in $\mathcal{O}_1$ are facts in the direct starting theory and, hence, they form an outlier set.

This piece of knowledge states that "*when the WK is on a edge rank and the distance between the rank of the WK and the rank of the BK is identical to the distance between the file of the WK and the file of BK (that is to say, both the Kings are positioned on the same diagonal), then it is the case that there not exists a strategy for the White player leading to a win in less the six moves*". The only exception to this rule is represented by the five examples belonging to the anomalous set $\mathcal{O}_1$. These examples concern the situation in which the WK and the BK are, respectively, in c1 and a3, and the RK is on the 4th rank. They state that in order for the White to have a winner strategy, the RK must stay on the file D or on a subsequent file (in these cases the White can win in exactly five moves). Indeed, if the RK were on file C, then White would not win in less than 11 moves. As for the RK on files A and B, the game terminates with a draw. By considering the full database, the above rule covers 1,129 example associated either with wins in more than five moves or draws.

The set $\mathcal{O}_2 = \{krk\_sl(c,1,a,8,a,1)\}$ is positive anomalous. Consider the associated dual starting theory:

$$not\_krk\_sl(A,B,C,D,C,E) \leftarrow edge(B), edge(D) \quad \langle 93 \rangle.$$

This rule states that "*when the WR and the BK lie on the same file and at the same time both the WK and the WR are on a (possibly different) edge rank, it is the case that the White cannot win within five moves*". As a matter of fact, c1-a8-a1 (a short for the WK-WR-BK configuration) is the only exception to such a rule (corresponding to a checkmate). The gain in this case is 0.0358.

In the full database there are 345 configurations captured by the above rule (all associated either with a depth of win greater than 7 or to a draw). E.g., consider the very similar configuration c1-a8-a2 covered by the same rule. In this case the White wins in 11 moves. Moreover, by relaxing the constraint that the WR and the BK lie on the same file, the knowledge is no longer valid. E.g., consider the close configuration c1-c8-a1, corresponding to a win for the White in only 2 moves.

The positive example set $\mathcal{O}_3 = \{krk\_sl(c,1,h,1,a,1), krk\_sl(d,1,h,1,a,1), krk\_sl(d,1,h,1,b,1)\}$ is anomalous. The associated dual starting theory is:

$$not\_krk\_sl(A,B,C,D,E,D) \leftarrow edge(B), edge(C) \quad \langle 45 \rangle$$
$$not\_krk\_sl(A,B,C,B,D,E) \leftarrow edge(B), edge(C) \quad \langle 28 \rangle.$$

The first rule says that "*if the WR and the BK are on the same rank, the WK is on an edge rank and the WR is on an edge file, then it is not the case that the White wins within five moves*". As for the second rule, it states that "*if the WR is in a corner of the chessboard and the WK is on the same rank, then the White cannot win within five moves*". The dual ending theory consists of 16 facts, and the gain is 0.013.

Note that in the whole database the two rules cover 345 and 372 examples respectively, concerning either wins in at least six moves or draws. By observing the distribution of the number of moves to win of the examples covered by the second rule, it appears that endgames satisfying the condition there stated are likely to exhibit a large number of moves to win. E.g., consider the configuration c1-h1-a2, very close to the configurations in the set $\mathcal{O}_3$ (the closer configuration there is c1-h1-a1 that leads to a win in 2 moves). It corresponds to a win in 12 moves.

The negative example set $\mathcal{O}_4 = \{krk\_sl(b,1,c,7,h,7), krk\_sl(c,1,c,6,h,6), krk\_sl(c,1,c,7,h,6), krk\_sl(c,1,c,7,h,7)\}$ is anomalous. The associated direct ending theory is

$$krk\_sl(A,B,C,D,E,F) \leftarrow edge(E), diff(A,E,G), diff(F,H,G) \quad \langle 552 \rangle.$$

This rule covers about the fifty percent of the positive examples. The direct starting theory contains a lot of facts and the gain 0.500. The associated knowledge can be stated as follows: "*Consider the situation in which the BK is on an edge file, and let Dist denote the distance between the file of the WK and the file of the BK. If the rank of the BK is such that there exists on the chessboard a rank at distance exactly Dist, then it is the case that the either the game ends in a draw or the White wins in no more than fourteen moves*". The above knowledge appears to be quite involved. However, it can

Noise (ν)	Abnormals removed	Positive Accuracy	Negative Accuracy	Total Accuracy
2.5%	0	93.04%	99.36%	97.26%
	10	95.81%	99.61%	98.35%
	20	97.08%	99.66%	98.80%
5%	0	82.00%	99.29%	93.38%
	10	87.06%	99.44%	95.21%
	20	89.90%	99.48%	96.21%
7.5%	0	71.87%	98.71%	89.30%
	10	78.94%	98.81%	91.85%
	20	81.26%	98.97%	92.76%
10%	0	64.43%	98.39%	86.19%
	10	68.36%	98.44%	87.63%
	20	69.55%	98.55%	88.13%

Table 3: Test accuracy for the *Illegal* data set in presence of noise after having removed abnormal sets.

it can be readily and more conveniently rephrased in the next one: "*if the BK is on and edge file and (at least) one of the two diagonals starting at its square intersects the file of the WK, then it is not the case that the White wins in more than fourteen moves*". The only exceptions to this rule are the four examples belonging to the set $\mathcal{O}_4$. It covers 7,706 database examples (about the 27.5% of the total).

7.3.4 Behavior in presence of noise

We wish to stress that the purpose of the approach here introduced is to gain domain understanding, and that it is not our explicit intention to regard abnormal sets as a tool for improving generalization.

Clearly, individuals showing an abnormal behavior may prevent the induction of a compact hypothesis, a condition which in its turn can negatively affect accuracy. Thus, one may wonder whether removing abnormal sets has some impact on accuracy or not.

For the *Illegal* data set a separate test set is available, composed of 3,361 positive examples and 6,639 negative examples (for a total of 10,000 test examples). The accuracy on the test set of the hypothesis induced on the training set is close to 100%.

In order to simulate the presence of noise, we flipped at random the sign of a controlled fraction ν of examples ($\nu \in \{2.5\%, 5\%, 7.5\%, 10\%\}$) belonging to the training set. We then measured the accuracy on the test set of the hypothesis induced both on the full noisy training set and on the noisy training set without negative abnormal examples returned by aCBOut for $k_{max} = 5$ and $\alpha = 0.01$.

Table 3 reports the results of the experiment. On the first column there is the noise level ν, while on the second column there is the number of abnormal sets removed from the set of examples. On the subsequent columns, the test

accuracy on the positive examples, the negative examples, and the whole set of examples is reported.

As expected, the test accuracy is directly related to the level of noise addition. Test accuracy benefits from the removal of abnormal sets. This circumstance is mainly evident on the positive class, which in this case is the class whose test accuracy worsens mostly by injecting noise. On the positive class, accuracy improves approximatively from four to ten percentage points. The increase in accuracy is maximum for the intermediate values of noise level here considered. This can be explained by noticing that the increase of accuracy depends on some opposing factors, that are the absolute accuracy, the level of noise addition, and the relative (w.r.t. the number of noisy examples) number of abnormal examples removed.

8 Conclusions

In this paper, a novel definition of outlier in the context of concept learning and effective techniques for singling them out have been presented. Our novel approach is designed for scenarios where there are no examples of normal or abnormal behavior, hence it is an unsupervised one, even if it has connections with supervised learning, since it is based on induction from examples.

Importantly, this approach is intended to provide a contribution in the framework of exploiting domain knowledge in order to improve the process of detecting outliers. As a matter of fact, most of the techniques presented in the literature for mining outliers are not able to take advantage of domain information, while it is clear that being able to incorporate a possibly available formal description of the domain of interest, e.g. encoded by means of a logic program, could great improve the quality of the process of outlier discovery. This direction of research has been only limited explored till now, and we have pointed out important differences with some techniques related to our one.

A further peculiarity of the introduced approach is to provide a finer characterization of the anomaly at hand, since we are able to distinguishing among three kinds of abnormalities: irregular, anomalous and outlier observations.

Other than learning more subtle forms of anomalies, we provide explanations for the detected abnormalities in the form of a pair of logic programs which make intelligible the motivation underlying their exceptionality.

As far as the applicability of our approach, both an exact and two approximate algorithms to mine abnormalities have been presented. The approximate algorithms improve execution time with respect to the exact algorithm while guaranteeing good accuracy. Experimental results confirmed the effectiveness of the proposed mining technique.

References

Aggarwal CC, Yu PS (2001) Outlier detection for high dimensional data. In: Proceedings of the International Conference on Management of Data (SIG-

MOD), pp 37–46
Angiulli F, Fassetti F (2009a) Dolphin: an efficient algorithm for mining distance-based outliers in very large datasets. ACM Transactions on Knowledge Discovery from Data (TKDD) 3(1)
Angiulli F, Fassetti F (2009b) Outlier detection using inductive logic programming. In: ICDM, pp 693–698
Angiulli F, Pizzuti C (2002) Fast outlier detection in large high-dimensional data sets. In: Proceedings of the International Conference on Principles of Data Mining and Knowledge Discovery (PKDD), pp 15–26
Angiulli F, Pizzuti C (2005) Outlier mining in large high-dimensional data sets. IEEE Trans Knowl Data Eng pp 203–215
Angiulli F, Basta S, Pizzuti C (2006) Distance-based detection and prediction of outliers. IEEE Trans Knowl Data Eng 18(2):145–160
Angiulli F, Greco G, Palopoli L (2007) Outlier detection by logic programming. ACM Trans Comput Log 9(1)
Angiulli F, Ben-Eliyahu-Zohary R, Palopoli L (2008) Outlier detection using default reasoning. Artif Intell 172(16-17):1837–1872
Bain M, Srinivasan A (1995) Inductive logic programming with large-scale unstructured data. Machine Intelligence 14
Breunig MM, Kriegel H, Ng RT, Sander J (2000) Lof: Identifying density-based local outliers. In: Proc. of the Int. Conf. on Manag. of Data (SIGMOD), pp 93–104
Bruno G, Garza P, Quintarelli E, Rosato R (2007) Anomaly detection through quasi-functional dependency analysis. Journal of Digital Information Management 5(4):190–200
Chandola V, Banerjee A, Kumar V (2009) Anomaly detection: A survey. ACM Comput Surv 41(3)
Chawla N, Japkowicz N, Kotcz A (2004) Editorial: special issue on learning from imbalanced data sets. SIGKDD Explorations 6(1):1–6
Debnath A, de Compadre RL, Debnath G, Shusterman A, Hansch C (1991) The structure-activity relationship of mutagenic aromatic nitro compounds. correlation with molecular orbital energies and hydrophobicity. Journal of Medicinal Chemistry 34:786–797
Fassetti F, Fazzinga B (2007) Approximate functional dependencies for xml data. In: ADBIS Research Communications, pp 86–95
He Z, Xu X, Huang J, Deng S (2005) Fp-outlier: Frequent pattern based outlier detection. Comput Sci Inf Syst 2(1):103–118
Hodge V, Austin J (2004) A survey of outlier detection methodologies. Artificial Intelligence Review pp 85–126
Kirsten M, Wrobel S, Horváth T (2001) Distance based approaches to relational learning and clustering. In: Džeroski S, Lavrač N (eds) Relational Data Mining, Springer, Germany, pp 213–232
Kivinen J, Mannila H (1995) Approximate inference of functional dependencies from relations. TCS 149:129–149
Knorr E, Ng R (1998) Algorithms for mining distance-based outliers in large datasets. In: Proc. of the Int. Conf. on Very Large Data Bases (VLDB), pp

392–403
Kriegel HP, Schubert M, Zimek A (2008) Angle-based outlier detection in high-dimensional data. In: KDD, pp 444–452
Lavrač N, Džeroski S (1994) Inductive Logic Programming, Techniques and Applications. Ellis Horwood
Lavrač N, Džeroski S, Bratko I (1996) Handling imperfect data in inductive logic programming. In: Raedt LD (ed) Advances in Inductive Logic Programming, pp 48–64
Liu FT, Ting KM, Zhou ZH (2012) Isolation-based anomaly detection. TKDD 6(1):3
Lloyd JW (1987) Foundations of Logic Programming. Springer-Verlag, Berlin
Mannila H, Räihä K (1987) Dependency inference. In: VLDB, pp 155–158
Muggleton S (1995) Inverse entailment and Progol. New Generation Computing, Special issue on Inductive Logic Programming 13(3-4):245–286
Muggleton S, Feng C (1990) Efficient induction of logic programs. In: First Conference on Algorithmic Learning Theory, pp 368–381
Muggleton S, Bain M, Hayes-Michie J, Michie D (1989) An exeperimental comparison of human and machine learning formalisms. In: Sixth International Workshop on Machine Learning
Novelli N, Cicchetti R (2001) Functional and embedded dependency inference: a data mining point of view. IS 26(7):477–506
Papadimitriou S, Kitagawa H, Gibbons PB, Faloutsos C (2003) Loci: Fast outlier detection using the local correlation integral. In: Proceedings of the International Conference on Data Engineering (ICDE), pp 315–326
Plotkin G (1971) A further note on inductive generalization, Machine Learning, vol 6, American Elsevier, chap 8, pp 101–124
Quinlan J, Cameron-Jones R (1993) Foil: A midterm report. In: 6th European Conference on Machine Learning, pp 3–20
Ramaswamy S, Rastogi R, Shim K (2000) Efficient algorithms for mining outliers from large data sets. In: Proceedings of the International Conference on Management of Data (SIGMOD), pp 427–438
Schölkopf B, Burges C, Vapnik V (1995) Extracting support data for a given task. In: KDD, pp 252–257
Srinivasan A, Muggleton S, Sternberg M, King R (1996) Theories for mutagenicity: A study in first-order and feature-based induction. Artif Intell 85(1-2):277–299