Consiglio Nazionale delle Ricerche, Istituto di Calcolo e Reti ad Alte Prestazioni (ICAR)
– Sede di Cosenza, Via P. Bucci 41C, 87036 Rende, Italy, URL: *www.icar.cnr.it*
– Sezione di Napoli, Via P. Castellino 111, 80131 Napoli, URL: *www.na.icar.cnr.it*
– Sezione di Palermo, Viale delle Scienze, 90128 Palermo, URL: *www.pa.icar.cnr.it*

Consiglio Nazionale delle Ricerche
Istituto di Calcolo e Reti ad Alte Prestazioni

High Quality True-Positive Prediction for Fiscal Fraud Detection

Stefano Basta[1], Fabio Fassetti[1], Gianfilippo Papi[2],
Stefano Pisani[3], Laura Spinsanti[4], Maurizio Atzori[4],
Fosca Giannotti[4], Massimo Guarascio[1],
Giuseppe Manco[1], Andrea Mazzoni[4], Dino Pedreschi[5].

1. ICAR CNR
2. SOGEI SpA
3. Agenzia delle Entrate
4. ISTI CNR
5. Università di Pisa

Rapporto Tecnico N.: **RT-ICAR-CS-09-1**

Data: **Gennaio 2009**

I rapporti tecnici dell'ICAR-CNR sono pubblicati dall'Istituto di Calcolo e Reti ad Alte Prestazioni del Consiglio Nazionale delle Ricerche. Tali rapporti, approntati sotto l'esclusiva responsabilità scientifica degli autori, descrivono attività di ricerca del personale e dei collaboratori dell'ICAR, in alcuni casi in un formato preliminare prima della pubblicazione definitiva in altra sede.

High Quality True-Positive Prediction for Fiscal Fraud Detection

Stefano Basta
ICAR-CNR
Cosenza, Italy
basta@icar.cnr.it

Fabio Fassetti
ICAR-CNR
Cosenza, Italy
f.fassetti@deis.unical.it

Gianfilippo Papi
SOGEI S.p.A.
Rome, Italy
gpapi@sogei.it

Stefano Pisani
Agenzia delle Entrate
Rome, Italy
stefano.pisani@agenziaentrate.it

Laura Spinsanti
KDDLab ISTI-CNR
Pisa, Italy
spinsanti@isti.cnr.it

ABSTRACT

Planning adequate audit strategies is a key success factor in *a posteriori* fraud detection applications, such as in fiscal and insurance domains, where audits are intended to detect fraudulent behavior. In this paper we describe an experience resulting from the collaboration among Data Mining researchers, domain experts of the Italian Revenue Agency, and IT professionals, aimed at detecting fraudulent VAT credit claims. The outcome is an auditing methodology based on a rule-based system, which is capable of trading among conflicting issues, such as maximizing audit benefits, minimizing false positive audit predictions, or deterring probable upcoming frauds. We describe the methodology in detail, and illustrate its practical effectiveness compared to classical predictive systems from the literature.

1. INTRODUCTION AND CONTEXT

Fraud detection represents a challenging issue in several application scenarios, and the automatic discovery of fraudulent behavior is a very important task with great impact in many real-life situations. In this context, fiscal fraud detection has witnessed an increasing interest and has become a widespread application field for data mining techniques.

In this paper we describe the experience we made on the Value Added Tax (VAT) fraud detection scenario. Like any tax, the VAT is open to fraud and evasion. There are several ways in which it can be abused, e.g. by underdeclaring sales or overdeclaring purchases. However, opportunities and incentives to fraud are provided by the credit mechanism which characterizes VAT: tax charged by a seller is available to the buyer as a credit against their liability on their own sales and, if in excess of the output tax due, refunded to them [14]. Thus, fraudulent claims for credit and refunds are an extensive and problematic issue in fiscal fraud detection.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
KDD '09 Paris, France
Copyright 200X ACM X-XXXXX-XX-X/XX/XX ...$5.00.

For example, [14] reports that 44 percent of all VAT fraud found in an investigation in the Netherlands took the form of false claims for tax paid at previous stages, for example by presenting forged invoices for non-existent or exaggerated purchases. The situation is further exacerbated in Italy by current laws which allows to compensate VAT credit with other taxes, thus boosting the trend for fraudulent behavior.

The *DIVA* project, that we report in this paper, tries to tackle the VAT Fraud Detection issue raised by the credit mechanism via the adoption of data mining techniques. The project involved computer science researchers, as well as experts from the Italian Revenue Agency and IT professional with expertise in managing the tax information system on behalf of the Italian Tax Administration. The objective of the project was to design a predictive analysis tool able to identify the tax payers with the highest probability of being VAT defrauders to the aim of supporting the activity of planning and performing effective fiscal audits. The construction of the model is based on historical VAT declaration records labeled with the outcome of the audit performed by the Agency.

The domain of the DIVA project is particularly challenging both from a scientific and a practical point of view. First of all, audited data available are only 0,004% of the overall population of taxpayers who file a VAT refund request. This resource-aware restriction inevitably raises a *sample selection bias*. Indeed, auditing is the only way to produce a training set, and auditors focus only upon subjects which are particularly suspicious according to some clues. As a consequence, the number of positive subjects (individuals which are actually defrauders) is much larger than the number of negative (i.e., non-defrauders) subjects. This implies that, despite the number of fraudulent individuals is far smaller than those of non-fraudulent individuals in the overall population, this proportion is reversed in the training set.

Since auditing is resource-consuming, the number of individuals reported as possible fraudsters is of high practical impact. Hence, a scoring system should primarily suggest subjects with a high fraudulent likelihood, while minimizing false positives. From a socio-economic point of view, it is preferable to adopt a rule based approach to modeling. Indeed, intelligible explanations about the reason why individuals are scored as fraudulent are by far more important than the simple scores associated to them, since they al-

low auditors to thoroughly investigate the behavioral mechanisms behind a fraud. Rule-based classifiers [20, 19, 11] are a mainstay of research in the field of concept learning, because of various desirable properties such as, e.g., their high expressiveness and understandability. Unfortunately, like most classification models, rule-based classifiers exhibit a poor predictive accuracy in highly imprecise learning settings, such as fraud and intrusion detection, where the underlying data distribution is inherently characterized by rarity and, hence, primary aspects of the concept to learn are infrequently observed. The situation is further exacerbated by the quest for a multi-purpose modeling methodology: typically, several objective functions characterize the fraud detection scenario, and a traditional classification scheme may fail in accomplishing such a multi-purpose task.

Former works in fraud detection rely on statistical techniques such as linear discriminant analysis or logistic discrimination [5]. Approaches based on the estimation of the underlying distribution and a direct modeling of the fraudulent behavior [4] exhibit shortcomings due to both the complexity of the domain under consideration, and the presence of noise which prevents suitable model fitting. Supervised techniques (where the a training set of already known fraudulent cases is available) typically suffer from the class-unbalance problem [9]: Typically, only a limited number of behavioral cases is recognized as fraudulent within the training set. Approaches based on hybrid or cost-sensitive classification techniques have been proposed and found effective in this context [8, 18, 12]. However, these techniques in general suffer from low interpretability which make them inadequate for the problem at hand. More generally, the above mentioned sample selection bias problem makes difficult to devise a proper training set upon which to rely.

Recently, semi-supervised techniques [21, 16, 10] have been proposed to partially overcome the drawbacks due to supervised techniques. In semi-supervised approaches, the training set of known cases is supported by further (unknown) cases, which can ease the learning process by refining the detection of the decision boundaries characterizing fraudulent behavior. Unsupervised techniques [3, 15, 7, 17, 1] do not need supervised information, and fraudsters are typically identified as outliers: the goal of finding outliers in a given data set is pursued by computing a score for each object suited to reflect its degree of abnormality. These techniques are quite effective in scenarios like anomaly-based intrusion detection, where real-time response is crucial. However, they again fail in providing interpretable explanations of the outlierness of a fraudster, although some initial study has been started in this context [2].

The contribution of this paper is the design of a supervised methodology capable of coping with all the above mentioned issues in a unified framework. The outcome that we describe in this work is *Sniper*, a predictive system capable of producing high-quality classification rules. On the basis of specific auditing requirements, the extracted rules in Sniper are able to select individuals from a population which likely exhibit high levels of proficiency, equity, and efficiency. It is worth to point out that although the above depicted issues are focused on the specific VAT refund fraud application case, similar concerns may actually arise on many real world situations. Clear enough, the framework presented in this paper goes beyond the specific application case and can be easily adapted whenever the illustrated aspects represent a challenging issue.

The paper is organized as follows. Section 2 introduces the main aspects of the problem addressed and of the Sniper technique, proposed to solve it. Section 3 formally describes the main problem tackled in this work and the subsequent Section 4 presents the proposed technique in details. Section 5 reports the description and the results of the application of Sniper on the real case study in the Italian scenario.

2. DIVA OVERVIEW

In this section we provide an overview of the experience we tackled and the related technique we propose. The section is intended to clarify the choices about the formal building raised up. The data coming from the governmental Revenue Agency is concerned with the VAT declarations of Italian business companies. In particular the experience is focused on the companies claiming a VAT refund. The data made available by the agency consisted of about 34 millions VAT declarations spread over 5 years. Data contain general 'demographic' information, like 'Zip of the registered office, 'start-up year' and 'Legal status', plus specific information about VAT declarations, like 'Business Volume', 'Sales', 'Import', 'Export' and the total amount of 'VAT Refund'. As a result of a data understanding process conducted jointly with domain experts, we chose a a total of 135 such features.

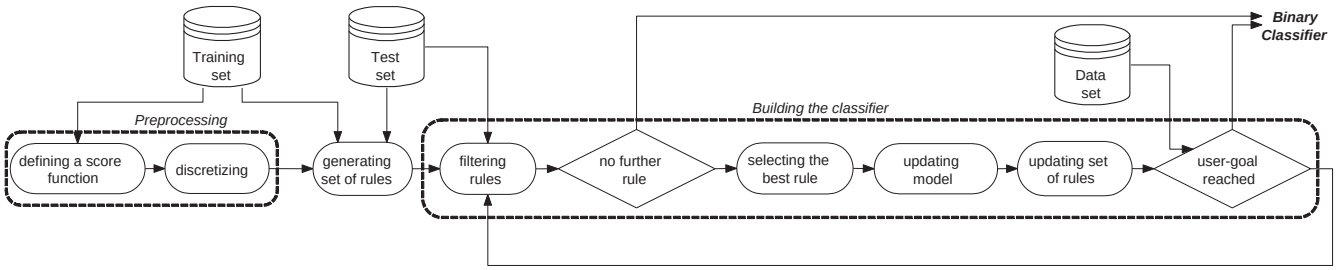
Out of the 34 millions declarations, we collected further information about 45,442 audited subjects. The results of auditing for such subjects are summarized in the further feature 'VAT refund fraud' (the difference between the amount of VAT Refund claimed and the VAT Refund actually due). Thus, audited subjects can be roughly classified into defrauders (when *'VAT refund fraud'* > 0) and non-defrauders (in the other case). The resulting labeled training set is extremely biased, consisting of 38,759(85.29%) subjects belonging to the "defrauder" class, and 6,683(14.71%) belonging to the "non-defrauder" class.

The situation is further exacerbated by the quest for a multi-purpose modeling methodology. Experts are interested in scoring individuals according to three main criteria:

Proficiency: scoring and detection should rely not only on a binary decision boundary separating defrauders from non-defrauders. Better, higher fraud amounts make defrauders more significant. For example, detecting a defrauder whose fraud amounts to 1,000$ is better than discovering a defrauder whose fraud amounts to 100$.

Equity: a weighting mechanism should leverage detection and scoring to include those cases where the amount of fraud is relevant related to their business volume. In practice, it should be avoided that individuals with low business volumes are never audited. For example, an individual whose fraud amounts to 1,000$ and whose business volume amounts to 100,000$, is less interesting than an individual whose fraud amounts to 1,000$ but the business volume amounts to 10,000$.

Efficiency: Since the focus is on refunds, scoring and detection should be sensitive to total/partial frauds. For example, a subject claiming an amount of VAT refund equal to 2,000 and entitled to 1,800 is less significant than a different subject claiming 200 and entitled to 0.



1: Flowchart of the SNIPER technique

A further requirement is represented by the limited auditing capacity of the Revenue Agency: auditing is a time-consuming task, involving several investigation and legal steps which ultimately require a full-time employ of human resources. As a consequence, the scoring system should retrieve from the population a user-defined fixed number of individuals with high defrauder likelihood.

SNIPER has been devised to accommodate all the above mentioned issues in a unified framework. The idea of the approach is to progressively learn a set of rules until all the above requirement are met. The approach is summarized in Figure 1.

As a first step, a scoring function is computed which associates an individual with a value representing its degree of interestingness according to the proficiency, equity and efficiency parameters. Clear enough, this function is not known for the individuals in the whole population. Nevertheless, the *training set* of audited subjects allows the computation of such a function and its analytical evaluation over those known cases.

A discretization step is accomplished for the scoring function, thus associating a class label to each discretization level. This leads to the definition of a class containing the individuals scoring to the maximum value of the function. Such a class is referred to in the following as *top class*.

The main objective is hence to build a rule set able to identify individuals belonging to the top class, with two main objectives: (i) false positives should be minimized; (ii) the number of subjects should be as close as possible to an user-specified value. To this purpose, a set of classifiers is trained, where each classifier provides a set of rules. These sets are collected in a global set $\mathcal{R}$ after a filtering phase that removes rules not complying with a minimum quality criteria. The set of rules $\mathcal{R}$ taken as a whole is not, in general, the best according to the two objectives cited above, since (i) its accuracy (the percentage of subjects of the top class retrieved) can be too low, and (ii) the number of retrieved subjects can be too high. This will be better clarified in Section 4.

The global set of rules $\mathcal{R}$ is employed as input in order to build a final binary classifier, consisting in the optimal subset of the rules in $\mathcal{R}$, according to the two main quality criteria. Notice that the problem of finding the best subset is intractable, thus triggering SNIPER to the adoption of a greedy strategy. The latter consists into iteratively selecting the "best" rule, until the quality criteria are met.

3. PROBLEM STATEMENT

This section provides a formal description of the different aspects which characterize the problem introduced in the previous section.

Some useful preliminary notions will follow. An *attribute* a is an identifier with an associated domain denoted as $Dom(a)$. Given a set of attributes $A = \{a_1, \ldots, a_h\}$, $Dom(A)$ denotes the set $Dom(a_1) \times \cdots \times Dom(a_h)$.

Let $A^d = \{a_1^d, \ldots, a_m^d\}$ and $A^c = \{a_1^c, \ldots, a_n^c\}$ be two sets of m and n attributes respectively, let $\mathcal{C}$ be a set of labels, and let $\perp$ is a special value standing for *unknown*. An *object* o on A^d, A^c and $\mathcal{C}$ is a triple $(\mathbf{v^d}, \mathbf{v^c}, c)$, where $\mathbf{v^d}$ is an m-ple $\langle v_1^d, \ldots, v_m^d \rangle$ of values, where $v_i^d \in Dom(a_i^d)$; $\mathbf{v^c}$ is an n-ple $\langle v_1^c, \ldots, v_n^c \rangle$ of values, where $v_i^c \in Dom(a_i^c) \cup \{\perp\}$; and $c \in \mathcal{C} \cup \{\perp\}$ is called the *class* of o. In the following $o[a_i^d]$ ($o[a_i^c]$, resp.) denotes the value v_i^d (v_i^c, resp.), while $class(o)$ denotes the label c.

A *dataset* D on two sets of attributes A^d and A^c and on a set of class label $\mathcal{C}$ is a multi-set of objects on A^d, A^c and $\mathcal{C}$. A^d is referred to as the set of *describing attributes*, namely the set of attributes which describe an object; while A^c is referred to as the set of *checking attributes*, namely the set of attributes whose value is known only for some objects and has to be predicted for the other objects. $\mathcal{C}$ is the set of class labels associated with the objects in D.

Let A be a set of attributes. A *condition* on A is an expression of the form $a \in V$, where $a \in A$ and $V \subseteq Dom(a)$. The expression $a \notin V$ is a shortcut for the condition $a \in Dom(a) \setminus V$.

Let D be a dataset on A^d, A^c, and $\mathcal{C}$. A *rule* on D is an expression of the form $B_0 \wedge \neg B_1 \wedge \cdots \wedge \neg B_k \rightarrow c$, where $B_0, \ldots, B_k$ are conjunction of conditions on A^d, and $c \in \mathcal{C}$. $B_0 \wedge \neg B_1 \wedge \cdots \wedge \neg B_k$ is called *body* of the rule, whereas c is the *head* of the rule; B_0 is the *positive* component of the rule, whereas $B_1, \ldots, B_k$ are the *negative* components of the rule. If the body of a rule is composed by only the positive component, then the rule is called *positive*.

For a rule $r : Body \rightarrow c$, $r.class$ denotes the class label c. A set of rules is also called *model*.

Given h rules, $r_1, \ldots, r_h$, they are said to be *same-head* rules if for each pair of rules r_i, r_j it holds that $r_i.class = r_j.class$. The size of a rule $r : B_0, \wedge \neg B_1 \cdots \wedge \neg B_k \rightarrow c$, denoted as $|r|$, is the number of conditions in B_0. An object o of D satisfies a conjunction $B = (a_1 \in V_1 \wedge \cdots \wedge a_m \in V_m)$ of m conditions if and only if $o[a_i] \in V_i, \forall i \in [1, m]$.

An object o of D is *activated* by a rule $r : B_0 \wedge \neg B_1 \wedge \cdots \wedge \neg B_k \rightarrow c$, if and only if o satisfies the positive component, B_0, and does not satisfy any negative component, $B_1, \ldots, B_k$, appearing in the body of r. The set of objects of D activated by a rule r is denoted as $r(D)$. The size of $r(D)$ is called *support* of the rule and denoted as $\sigma(r)$.

A rule r is *exclusive* with respect to a a rule r' on the dataset D, if no object in D activated by r is activated also by r', namely if $r(D) \cap r'(D) = \emptyset$.

The objects activated by a rule $r : Body \rightarrow c$ whose class is actually c are called *true positive*, the other objects activated by r are called *false positive*.

Definition 1. Let $r : Body \rightarrow c$ be a rule on a dataset D labeled w.r.t. a set of labels $\mathcal{C}$. The confidence of r, denoted as $\gamma(r)$ is the ratio between the true positive objects activated by r and the support of r.

The above notions, given for a single rule, can be naturally extended to a set of same-head rules. An object o is said to be activated by a set of same-head rules $\mathcal{R}$ if and only if it is activated by at least one rule $r \in \mathcal{R}$; more formally, the set of objects activated by a set of same-head rules $\mathcal{R}$ is $\mathcal{R}(D) = \bigcup_{r \in \mathcal{R}} r(D)$. A rule r is exclusive with respect to a set of same-head rules $\mathcal{R}$ if and only if $r(D) \cap \mathcal{R}(D) = \emptyset$. The support of a set of same-head rules is $\sigma(\mathcal{R}) = |\mathcal{R}(D)|$, while the confidence $\gamma(\mathcal{R})$ is the ratio between the true positive objects activated by $\mathcal{R}$ and the support of $\mathcal{R}$.

Finally, the $\bar{\wedge}$ operator is introduced. Let $r : B_0 \wedge \neg B_1 \wedge \cdots \wedge \neg B_k \rightarrow c$ be a rule and $r' : B'_0 \rightarrow c$ be a positive rule with the same head as r. $r \,\bar{\wedge}\, r'$ denotes the rule $r'' : B_0 \wedge \neg B_1 \wedge \cdots \wedge \neg B_k \wedge \neg B'_0 \rightarrow c$. Note that r'' is exclusive with respect to r'. In other words, $r \,\bar{\wedge}\, r'$ produces a rule that activates all the objects activated by r and not by r'.

As previously discussed, the main problem to be solved is to identify the individuals of a given population showing the most exceptional behavior; and additionally, the number of individuals to be retrieved is fixed by the user. Moreover, an explanation about the reason for which the retrieved individuals are detected as the most interesting ones should be provided to the user together with the individuals.

To formally define the problem, the meaning of "exceptional behavior" must be provided. The idea is to follow a requirements-based approach, where the analyst together with the user define an interesting function (called *score function* in the following) that measures the exceptionalness of each individual. The goal is, then, to retrieve the individuals scoring the maximum value of such interesting function. Such a function is assumed to be not evaluable on the individuals of the given population, and then the challenge is to predict its value. Nevertheless, the value of the score function is known for a certain set of individuals, referred to as *training set* in the following, which are the individuals selected by domain experts for being audited.

The second important issue is to provide an explanation intelligible for the user, justifying the exceptionalness of the retrieved individuals. The approach here pursued is to return a set of rules which directly provides an intelligible justification for the exceptionalness of an individual. For example, the rule r : Start-up Age $> 2 \wedge$ Sales $=$ *high* $\rightarrow$ *defrauder*, immediately provides the user of a semantically significant explanation about defrauders. In particular, r asserts that any company claiming a VAT refund while exhibiting more than two years of activity (Start-up Age > 2) and a high amount of Sales is a defrauder.

The main problem we aim to solve can be formally defined as follows.

Definition 2. Given a dataset D, a scoring function Ω, three thresholds $\sigma^{\min}$, $\gamma^{\min}$ and X, the problem is: *find the set $\mathcal{R}$ of rules each having at least confidence $\gamma^{\min}$ and support $\sigma^{\min}$ such that $|\mathcal{R}(D)|$ is as close to X as possible, and the objects in $\mathcal{R}(D)$ score the highest values of Ω.*

4. SNIPER TECHNIQUE

In this section the SNIPER technique, designed to solve the previously described problem, is presented. The main steps of the technique are summarized in Figure 1 and can be roughly subdivided into two main parts: the *preprocessing* phase and the phase concerning the *building of the classifier*.

4.1 Preprocessing

The main objective of this phase is to formalize the notion of interestingness and exceptionalness of an individual. As already stated, auditing individuals is a very resource-consumption task and then it should be focused on those individuals which, among the defrauders, are the most interesting ones Thus a scoring function capable of ranking the whole population (and of detecting the *top-fraudulent* individuals) is preferable to a rough classification of the population into fraudulent and non-fraudulent individuals.

The notion of "interesting" subjects is domain-dependent. In general, many aspects should be taken into account for dealing with the user notion of interesting individuals, and then many parameters contribute at identifying an individual as interesting. The idea here pursued is to define, together with the user, a function (called *first-level function*) for each of such parameters; then to combine them in a *second-level function*, able to weight the different first level functions in order to match as better as possible the user needs; and finally to define, starting from the second level function, a score function able to assign an interesting value to each training set individual.

4.1.1 First-level functions

Definition 3. Let D be a dataset on the sets of attributes A^d, A^c and $\mathcal{C}$. A *first-level objective function* on D is a function $f : Dom(S^d) \times Dom(S^c) \rightarrow \mathbb{R}$, where S^d is a possibly empty subset of A^d and S^c is a non-empty subset of A^c.

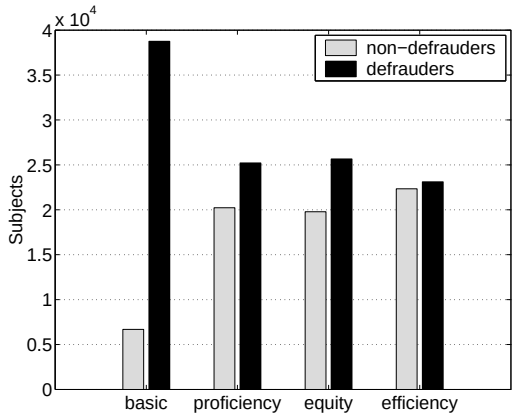
It follows from the definition that f requires the values of some attributes in A^c, then f can be evaluated on the individuals of the training set, but not on all the individuals in the whole population.

For each first-level objective function f, two thresholds σ_f^{up} and σ_f^{low} are defined with $\sigma_f^{up} \geq \sigma_f^{low}$. These thresholds split the codomain of f in three sets, and accordingly also the individuals can be split into three sets:

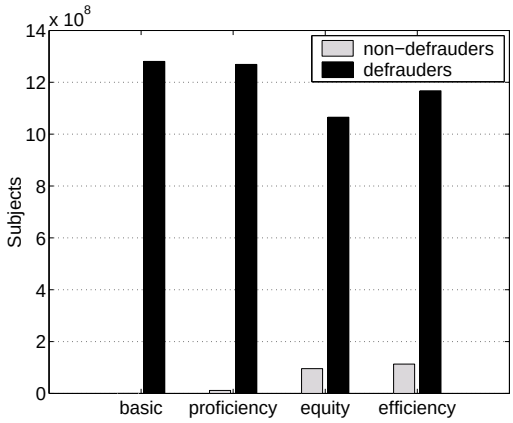
1. $S_f^{low} = \{o \in D \mid f(o) \in (-\infty, \sigma_f^{low})\}$
2. $S_f^{mid} = \{o \in D \mid f(o) \in [\sigma_f^{low}, \sigma_f^{up}]\}$
3. $S_f^{up} = \{o \in D \mid f(o) \in (\sigma_f^{up}, \infty)\}$

The above thresholds can also assume value $\pm\infty$, meaning that some of the above defined sets can collapse.

The importance of the thresholds is to let the score function be influenced by outstanding behaviors on a single first-level function. Indeed, an individual assuming a very high value on a first-level function f (beyond σ_f^{up}) can be of interest for the user even if its score (valuated on the combination of all the first-level functions) is not very high. Analogously,



2: Training set partitioning according to first-level functions



3: Retrieved fraud within the partitioned dataset.

if an individual that assumes a very low value on a first-level function (below σ^{low}) and it is not outstanding (beyond σ^{up}) in any of the first-level functions, then it is not interesting for the user even if its score is high.

In the VAT refund fraud contest, we can devise three first-level functions, which model the notions of proficiency, equity and efficiency described in section 2. In particular, f_{prof} represents the total amount of fraud, whereas f_{equ}, is defined as the ratio between the total fraud and the business volume, and f_{eff} is defined as the ratio between the total fraud amount and the total VAT refund declared. Each of these functions model a different notion of interestingness for the subjects in the training set. In particular, for each function f the subjects whose value assumed on f is higher than σ_f^{low} are classified as "interesting".

Figure 2 reports the distribution of defrauders and non-defrauders subjects belonging to the training set. The first histogram represents the distribution as partitioned by f_{prof} with threshold 0. Note that this corresponds to roughly classify as non-defrauders those subjects whose total fraud amount is 0 and as defrauders all the other subjects, denoted as f_{basic}. The other histograms represent the distribution as partitioned by f_{prof}, f_{equ}, and f_{eff}, respectively. The low thresholds employed for these functions, chosen with the aid of domain experts, are $\sigma^{low}_{f_{prof}} = 2{,}000$, $\sigma^{low}_{f_{equ}} = 0.0025$, and $\sigma^{low}_{f_{eff}} = 0.2$.

It is important to notice that a careful choice of the threshold values allows to alleviate the sample selection bias, and contemporarily does not alter the significance of the training set. Figure 3 shows the retrieved fraud (i.e., the sum of the VAT refund fraud) associated with both the subjects identified either as defrauders or as non-defrauders subjects by each of the first level functions above considered. Figure 2 highlights that the size of the set S_f of subjects identified as defrauders by the first level function f is strongly reduced with respect to the size of the set $S_{f_{basic}}$ of defrauders identified by f_{basic}. Nevertheless, as shown in figure 3 the retrieved fraud of S_f is almost similar to that in $S_{f_{basic}}$, thus confirming that the most interesting defrauders are those selected by the first-level functions.

4.1.2 Second-level functions

First-level functions play a major role in modeling local properties of fraudulent behavior. The role of a second-level function is to combine such local properties into a global interestingness measure capable of summarizing them.

More formally, given k first level objective functions $f_1, \ldots, f_k$, a *second-level objective function* is a function $\mathcal{F}$, associating each individual of a population with a real number ranging in [0,1], by combining the values assumed by $f_1, \ldots, f_k$. The contribution of f_i can also be weighted, in order to tune its influence within $\mathcal{F}$.

The combination is made of two step. A first preliminary step consists in harmonizing the values of the first-level functions. Indeed, first-level functions are designed independently to each other and to capture different features. Thus, often they are in different ranges and in different scales. Consider for example, the function f_{prof} and the function f_{equ}. The former represents the absolute value of the fraud amount, while the latter represents the ratio between the fraud amount and the business volume, thus ranging in $[0, 1]$. Directly combining them is clearly misleading as they refer to different unit measures.

Harmonization should also take care of rescaling values according to threshold values, in order to preserve homogeneity in comparisons. Consider for example two functions f_1 and f_2, both ranging in $[0, 1]$, whose thresholds are $\sigma_1^{low} = 0.01$, $\sigma_1^{up} = 0.1$, $\sigma_2^{low} = 0.7$ and $\sigma_2^{up} = 0.9$. If for an object o both $f_1(o)$ and $f_2(o)$ assume value 0.5, the semantic of such a value is inherently different, and a combination of such values without a proper adjustment would result into a misleading score.

Within SNIPER, harmonization is accomplished by means of a *normalizing* function $\mathcal{N} : \mathbb{R} \to [0, 1]$, associating each value assumed by a first level function with a value in the range $[0, 1]$. $\mathcal{N}$ can simultaneously account for the normalization concerning scales, ranges and thresholds. In DIVA we adopted hyperbolic functions for normalizing values and making them comparable.

Second-level functions can be directly derived by combining and weighting the normalized versions of the first-level functions. We considered two main combination functions:

$$\mathcal{F}_{\Pi}(o) = \prod_{i \in [1,k]} (\mathcal{N}(f_i(o)))^{p_i}$$

$$\mathcal{F}_{\Sigma}(o) = \sum_{i \in [1,k]} p_i \cdot \mathcal{N}(f_i(o)),$$

where p_i represents the weight associated with f_i. The $\mathcal{F}_\Pi$ function returns the weighted product of the f_i, whereas the $\mathcal{F}_\Sigma$ function returns the weighted sum of the f_i.

These two functions satisfy a different conceptual enforcement, but both of them have guaranteed good experimental results. Essentially, the former function is built by applying a sort of conjunctive operator to the single first-level functions; this fact causes that $\mathcal{F}_\Pi(o)$ assigns an higher value to those subjects having high values for each first level functions. The latter instead implements a disjunctive criteria, which associates a high value with those subjects having an high value for one first level functions at least.

Thus, the $\mathcal{F}_\Pi$ function is more selective than $\mathcal{F}_\Sigma$ and therefore it could assign a low value to some interesting subjects, for instance, characterized by a low value for one first level function at least and a very high value for the other ones. Analogously, $\mathcal{F}_\Sigma$ suffers of the opposite problem, namely, it could assign an high value to those subjects having an high value for one first level function but low values for all the other ones.

In principle, any second-level function can be used as a score function. We chose, however, to mitigate the effects on the borders by directly controlling them. Hence, the score function can be formally defined as follows.

Definition 4. Let D be a dataset, let $f_1, \ldots, f_k$ be k first level objective functions on D on which are defined $2k$ thresholds $\sigma^{low}_{f_1}, \sigma^{up}_{f_1}, \ldots, \sigma^{low}_{f_k}, \sigma^{up}_{f_k}$, let $p_1, \ldots, p_k$ be k weights, and let o be an object of D. A *score function*, or simply *score*, is a function Ω such that:

$$\Omega(o) = \begin{cases} 0 & \text{if } \bigvee_{i\in[1,k]} o \in S^{low}_{f_i} \wedge \bigwedge_{i\in[1,k]} o \notin S^{up}_{f_i} \\ \mathcal{F}(o) & \text{otherwise} \end{cases}$$

where $o \in D$ and $\mathcal{F}(\cdot)$ is a second level function defined on $f_1, \ldots, f_k$ and $p_1, \ldots, p_k$.

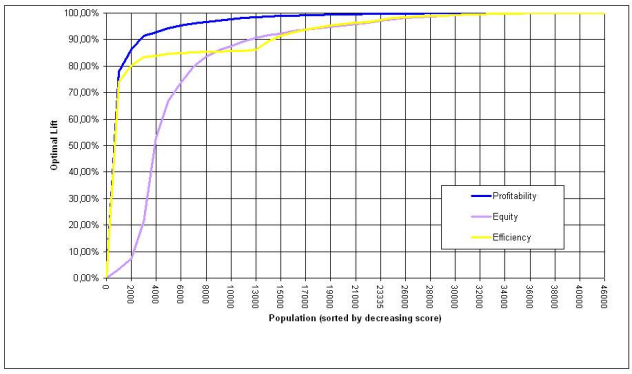
Then, the score of an object o evaluates 0 if o belongs to S^{low}_f for some first level function f, unless o belongs to $S^{up}_{f'}$ for some first level function f'; otherwise the score is the value of a suited second level-function assumed by the object.

The adequacy of Ω for capturing the most prominent aspects of the first-level functions can be appreciated in figure 4. Here, we show the cumulative gains obtained for decreasing values of the score function (equipped with $\mathcal{F}_\Pi$), relative to profitability, equity and efficiency. Notice that top individuals cumulate the largest gain in practically all the three parameters.

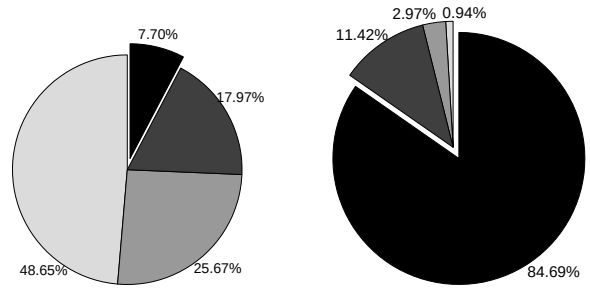
4.1.3 Discretization

In the framework we are addressing, the goal is to retrieve from the population X individuals scoring the maximum value for the score function. Since Ω is a continuous function, then the score function must be discretized, so that a class label can be associated with each discretization interval. Then, the interval containing the highest values of the score function identifies the top class. The width of the top class strongly influences the quantity and the quality of the individuals identified as members of the top class in the population. Hence, the discretization phase plays a very important role.

The approach used for the discretization in our experience is detailed next. First of all, two thresholds, ω^{low} and



4: Cumulative gains in proficiency, equity and efficiency related to the score function.



(a) Subject partitioning (b) Retrieved fraud

5: Score function results

ω^{up} are defined for the score function Ω. These thresholds are obtained by computing the value that the second level function assumes when evaluated on the normalized values of the first level function thresholds. Then, the subjects are partitioned in classes according to the threshold levels:e.g., the top class is made of subjects whose score is greater than ω^{up}.

Figure 5a reports the effects of the employed discretization in partitioning the subjects. Specifically, from the lighter to the darker colored slice, the figure reports the percentage of subjects in classes 0, 1, 2 and 3, respectively. Conversely, Figure 5b reports the percentage of total amount of fraud made by the subjects of the different classes. It is worth to point out that the subjects belonging to the top class, represent only 7.70% of the total number of audited subjects, but the total amount of fraud associated with them is 84.69% of the total fraud amount made by the whole set of audited subjects. This confirms the adequacy of the score function and the related discretization to our needs.

4.2 Building the classifier

The second part of the SNIPER technique is the building of the final binary classifier able to identify X defrauders in the dataset, likely to be the most fraudulent individuals. As introduced in Section 2, the SNIPER technique trains a set of classifiers on the top class detected in the training set by the preprocessing phase. This classifiers are then merged into a single ruleset which is further processed.

4.2.1 Generating rules

The first step consists in extracting the rules from the

training set. The adoption of a single classifier directly trained over the training is infeasible: as shown in fig. 4.1.3, the preprocessed training set is highly unbalanced w.r.t. the top class, and even the adoption of advanced mechanism for resampling or cost-sensitive learning produces low-accuracy models. The reason for this phenomenon is well-known in literature as the problem of *rare cases*. The latter are very small portions of the training data, that can be viewed as exceptional sub-concepts seldom occurring within predominant or rare classes. In the VAT refund fraud scenario, this corresponds to the fact that each defrauder has a peculiar behavior that does not generalize to other defrauders. As pointed out in [22], rarity actually prevents a rule-based classifier from finding and reliably generalizing the regularities within infrequent classes and exceptional cases. Indeed, due to the commonly adopted metrics for growing classification rules and evaluating their accuracy, class imbalance leads to several accurate rules targeting the predominant classes, supplemented by very few (if any) error-prone rules predicting minority classes. Furthermore, rare cases tend to materialize within the resulting classifier as strongly inaccurate rules, referred to as *small disjuncts* [13], that in most cases do not generalize actual exceptions, being a mere consequence of noise in the training data [23] . In highly imprecise learning settings, noise often contributes to the effects of rarity on predictive accuracy. On one hand, it may further skew the already unbalanced class distribution. On the other hand, rare cases may appear to the learner as indistinguishable from noise, thus requiring a more specific inductive bias, that would ultimately also induce noisy small disjuncts.

The solution provided by SNIPER consists in building a hybrid classifier, resulting from the combination of the whole set of classifiers trained over the training set. The approach is similar in spirit to a bagging methodology [6]. However, rather than implementing a voting mechanism over an independent set of similar rule-based models, we chose to decompose each classifier into a single ruleset and to merge all the rulesets into a global ruleset $\mathcal{R}$, from which to extract the most prominent rules.

Decoupling the model construction phase from model selection, provides us with the further advantage of approaching the *rare case* problem with a brute-force approach: in the model construction, several different strategies are attempted to build models specialized on local peculiarities of the top class. In the model selection phase, several local fragments can be combined or discarded if the global accuracy improves.

4.2.2 Merging rulesets

Let $R_1, \ldots, R_h$ be the set of rules returned by h classifiers, and let *top* be the class label assigned by the classifiers to the objects belonging to the top class. The candidate ruleset $\mathcal{R}$ is defined as follows:

$$\mathcal{R} = \left\{ r \in \bigcup_{i \in [1,h]} R_i \mid r.class = top \right\}$$

The ruleset $\mathcal{R}$ still represents a classifier, and class *top* is assigned to a non-labeled object o if and only if there exists at least a rule in $\mathcal{R}$ that activates it. Hence, all and only the objects in $\mathcal{R}(D)$ are labeled *top*.

Taken as a whole, the global ruleset $\mathcal{R}$ presents two relevant shortcomings. first, $|\mathcal{R}(D)|$ can be larger than X. Second, the confidence of $\mathcal{R}$ can be too low, and in particular, it could be lower than $\gamma_{\min}$. Indeed, $\mathcal{R}$ is the result of merging different and independently designed classifiers which are not necessarily exclusive.

Assume, for the sake of simplicity, that $\mathcal{R}$ is composed by only two rules r_1 and r_2 having confidence $\gamma(r_1) = \frac{p_1}{p_1+n_1}$ and $\gamma(r_2) = \frac{p_2}{p_2+n_2}$, respectively. Here, p_i (resp., n_i) denotes the number of true (resp., false) positive objects activated by the rule r_i.

Let $p_{1,2}$ (resp., $n_{1,2}$) be the number of true (resp., false) positive objects activated by both r_1 and r_2; with $p_{1,2}$ ranging in $[0, \min\{p_1, p_2\}]$ and $n_{1,2}$ ranging in $[0, \min\{n_1, n_2\}]$.

Then, the global confidence of $\mathcal{R} = \{r_1, r_2\}$ is:

$$\gamma(\mathcal{R}) = \frac{p_1 + p_2 - p_{1,2}}{p_1 + n_1 + p_2 + n_2 - p_{1,2} - n_{1,2}}.$$

Hence, the maximum value of $\gamma(\mathcal{R})$ is obtained when $p_{1,2} = 0$ and $n_{1,2} = \min\{n_1, n_2\}$,

$$\gamma_{\max}(\mathcal{R}) = \frac{p_1 + p_2}{p_1 + n_1 + p_2 + n_2 - \min\{n_1, n_2\}},$$

which is the case when the sets of true positive objects activated by r_1 and r_2 are disjoint, whereas the sets of false positive objects activated by r_1 and r_2 are overlapped.

Conversely, the minimum value of $\gamma(\mathcal{R})$ is obtained when $p_{1,2} = \min\{p_1, p_2\}$ and $n_{1,2} = 0$,

$$\gamma_{\min}(\mathcal{R}) = \frac{p_1 + p_2 - \min\{p_1, p_2\}}{p_1 + n_1 + p_2 + n_2 - \min\{p_1, p_2\}},$$

which is the case when the sets of true positive objects activated by r_1 and r_2 are overlapped, whereas the sets of false positive objects activated by r_1 and r_2 are disjoint.

It is worth to point out that $\gamma_{\max}(\mathcal{R})$ can be larger than $\max\{\gamma(r_1), \gamma(r_2)\}$, whereas $\gamma_{\min}(\mathcal{R})$ can be smaller than $\min\{\gamma(r_1), \gamma(r_2)\}$. This depends from both the confidences and the supports of the rules r_1 and r_2.

In case the rules are exclusive, both $p_{1,2}$ and $n_{1,2}$ are equal to 0. Then,

$$\gamma(\mathcal{R}) = \frac{p_1 + p_2}{p_1 + n_1 + p_2 + n_2}.$$

Suppose, w.l.o.g., that $\gamma(r_1) < \gamma(r_2)$, and then that $\frac{p_1}{p_1+n_1} < \frac{p_2}{p_2+n_2}$. It follows that:

$$\gamma(\mathcal{R}) = \frac{p_1 + p_2}{p_1 + n_1 + p_2 + n_2} > \frac{p_1 + p_2}{p_1 + n_1 + \frac{p_2}{p_1}(p_1 + n_1)} > \frac{p_1}{p_1 + n_1}$$

Analogously, it can be shown that $\gamma(\mathcal{R}) < \frac{p_2}{p_2+n_2}$.

Summarizing, if the rules are exclusive the value of the global confidence $\gamma(\mathcal{R})$ is greater than the minimum confidence of the rules in $\mathcal{R}$ and lower than the maximum confidence of the rules in $\mathcal{R}$; conversely, these properties do not hold if the rules are not exclusive.

Thus, $\mathcal{R}$ is not necessarily the optimal choice for the final binary classifier. We can, however, look for an optimal subset $\mathcal{R}^* \subset \mathcal{R}$, which simultaneously reaches the two following goals: (i) the number of objects retrieved in the dataset is as close to X as possible, and (ii) the confidence of $\mathcal{R}^*$ is as high as possible.

The search for the best subset $\mathcal{R}^*$ achieving these two goals is referred to as SBR problem in the following. Solving such a problem is a hard task. Unfortunately, the SBR problem can be proved to be NP-hard.

The complexity of the SBR problem is formally provided in the appendix.

Input: A set of non-exclusive positive rules $\mathcal{R}$, a confidence threshold $\gamma_{\min}$, an integer X
Output: A model $\mathcal{M}$
Method:
1: $\mathcal{M} := \emptyset$
2: $\mathcal{R} := \left\{ r \in \mathcal{R} \mid \gamma(r) \geq \gamma_{\min} \right\}$
3: **while** $\mathcal{R} \neq \emptyset$ **do** *//first stop condition*
4: $r^* := \arg\max_{r \in \mathcal{R}} \{\gamma(r)\}$ *//select the best rule*
5: $\mathcal{M} := \mathcal{M} \cup \{r^*\}$ *//update the current model*
6: **if** $\mathcal{M}(D) \geq X$ **then** *//second stop condition*
7: **return** $\mathcal{M}$
//update the set of rules
8: $\mathcal{R} := \left\{ r' = r \barwedge r^* \mid (r \in \mathcal{R} \setminus \{r^*\}) \bigwedge (\gamma(r') \geq \gamma_{\min}) \right\}$
9: **return** $\mathcal{M}$

6: Selecting Best Rules Algorithm

4.2.3 Selecting the best rules

In this section we describe a greedy technique for obtaining the resulting ruleset, starting from $\mathcal{R}$. Loosely speaking, the heuristic employed consists in iteratively taking the most confident rules until X objects are retrieved from D, or until no further rules with enough confidence exist in $\mathcal{R}$.

The algorithm is shown in Figure 6. We employ the term *set of rules* to refer to the input set of same-head rules coming from the classifiers, whereas the term *model* refers to the set of rules finally computed by the algorithm. The main idea is to compute a model $\mathcal{M}$ by iteratively adding the most confident rule to it. Since rules may overlap, the confidence of the rules is evaluated with regards to the objects not activated by the model $\mathcal{M}$ associated with the current iteration, rather than the whole test set.

First of all, the algorithm removes from the input set $\mathcal{R}$ those rules that are not at least $\gamma_{\min}$ confident. Then, the most confident rule r^* in $\mathcal{R}$ is selected and added to $\mathcal{M}$ (*lines* 4-5). Next, the set $\mathcal{R}$ is updated by removing r^* and by replacing each rule r other than r^* with the rule $r' = r \barwedge r^*$ if $\gamma(r') = \gamma_{\min}$, otherwise r is just removed from $\mathcal{R}$ (*line* 8).

In such a way, the rules which are now in $\mathcal{R}$ can activate only objects which are not contemporarily activated by any other rule in $\mathcal{M}$. In other words, each rule in $\mathcal{R}$ is exclusive with respect to the set $\mathcal{M}$ of rules.

The main property of the algorithm consists in the fact that, for a given rule r, the more the set of true positives activated overlaps with the true positives activated by $\mathcal{M}$, the higher the confidence of $r' = r \barwedge r^*$ is. Hence, each iteration selects the rule that gives the best contribution to the global confidence of the model $\mathcal{M}$. Moreover, since at each iteration adds to $\mathcal{M}$ a rule which is exclusive with respect to $\mathcal{M}$, and since the confidence of such a rule is at least $\gamma_{\min}$, the confidence of $\mathcal{M}$ cannot be lower than $\gamma_{\min}$. This guarantees that the heuristic produces a high-quality model.

The algorithm proceeds until one of the two stopping conditions is reached, namely either no other rule is in $\mathcal{R}$ (*line* 3), or $\mathcal{M}$ activates X objects in the dataset (*line* 6).

5. RESULTS

In this section, we briefly show the main experimental results obtained applying the SNIPER technique to the real-life VAT refund fraud scenario so far considered.

5.1 Learning of single classifiers

First of all, we have separately computed several classification models using a score function to label the examples belonging to training set. Precisely, we have preferred to exploit the score function based on the $\mathcal{F}_\Pi$ function in order to better fit the domain's constraints. The classifiers have been selected from the Weka workbench [24] and other commercial tools. Several different parameters sets were adopted, including cost models for cost-sensitive learning. The results of some experiments are reported in Table 1. The experiments marked with a "$*$" refer to classifiers modified in order to improve their performance in terms of subjects retrieved. That is, if the underlying original classifier extracts more than X subjects, the less confident rules are removed until a number of subjects close to X is retrieved from the dataset. Note that since all the algorithms employed extract a model with exclusive rules, if the less confident rule is removed from the model, the global confidence raises up.

For each classifier C_i, the table contains information about the support and the confidence of the model extracted by C_i on the test set (*columns 2-3*); and finally, the number of subjects of the dataset identified as fraudulent by C_i. The classifiers are ordered by increasing value of confidence.

None of the single classifiers satisfies our quality needs. Indeed, they are not able to simultaneously ensure a small number of false positives and a number of dataset subjects retrieved close to $X = 10{,}000$. In particular, high-quality models are only capable of selecting a small number of subjects from the whole population, which is too far from the value X required. Conversely, larger auditing sets can only be obtained by low-accuracy classifiers.

classifier ID	*supp (%)*	*conf (%)*	*dataset subjects*
C_1	1.01	84.90	1,910
C_2	1.10	82.97	2,240
C_3	3.11	77.28	4,955
C_4	3.44	77.12	5,675
C_5^*	6.36	62.26	10,056
C_6^*	6.81	60.80	8,875
C_7^*	7.07	59.72	9,059
C_8^*	5.22	52.64	9,950
C_9^*	4.56	49.18	12,584

1: Single classifiers behavior

5.2 Sniper technique results

rule	*1*	*2*	*3*	*4*	*5*	*6*	*7*
supp	0.65	1.21	0.97	0.89	0.85	0.87	0.90
conf	97.64	94.53	88.41	88.09	87.76	87.66	87.29
rule	*8*	*9*	*10*	*11*	*12*	*13*	*14*
supp	1.01	0.12	0.17	0.52	0.26	0.17	0.19
conf	85.12	83.64	83.12	77.73	76.47	71.79	70.11

2: Rules of the final classifier

By contrast, the SNIPER technique steps until a small set of rules, extracted from the set containing all the rules of each classifiers, comes out as the final result of the procedure. The parameters we adopt in the experiments are $\sigma_{\min} = 0.1\%$ (corresponding to 50 subjects), $\gamma_{\min} = 70\%$, and $X = 10{,}000$.

Notice that, according to described approach, the mechanism governing rules' selection guarantees that the returned set of rules is characterized by a global confidence value

greater than the fixed threshold $\gamma_{\min} = 70\%$. Precisely, the final model contains 14 rules coming from 9 distinct classifiers. Table 2 shows the characteristics of each of these rules. The global confidence of the final model $\mathcal{M}$ is $\gamma(\mathcal{R}) = 80.41\%$; whereas the number of total subjects of the dataset it activates is 9,840. Such results witness how the SNIPER technique outperforms single classifiers.

6. CONCLUSION AND DISCUSSION

In this paper we presented SNIPER, a predictive modeling technique for multi-purpose fiscal fraud detection in presence of biased and unbalanced training sets. The methodology produces a rule-based classification system that can be tuned to the requirements of the auditing agency, since it concentrates on a user-defined fixed number of most prominent subjects recognizable as fraudsters. The methodology has been applied, in collaboration with the Italian Revenue Agency, to the case of VAT refund fraud, although it can be generalized to other situations where fraud detection can be characterized by a multi-purpose objective in presence of a noisy environment.

The SNIPER methodology is currently being validated on stage: a number of subjects have been selected on the basis of the SNIPER rules, and actual audits are being performed, in order to asses the predictive accuracy and effectiveness. At the time this paper is written, we can report only some preliminary results

Two relevant outcomes are currently substantiating in the validation process. The first is that most of the audited subjects are unexpected cases, i.e., subjects the experts should have never selected for auditing based on their current practices. That is, the adoption of Data Mining methodology can ease the discovery of new fraud behaviors. The second result is that audited subjects found positive typically met all the three criteria of proficiency, equity and efficiency. Proficiency and efficiency exhibit values close to those in the top class in the training set. Equity, by contrast, exhibit an impressive higher values, with increases ranging from 1% to 37%. The meaning is that the model succeeds in pursuing a multi-purpose objective, being in particular able to identify subjects with high fraud with respect to business volume. Since these subjects were generally ignored by current audit practices, the SNIPER methodology may represent a significant advance in strategic planning for fiscal fraud detection.

7. ADDITIONAL AUTHORS

Maurizio Atzori (ISTI-CNR), Fosca Giannotti (ISTI-CNR), Massimo Guarascio (ICAR-CNR), Giuseppe Manco (ICAR-CNR), Andrea Mazzoni (ISTI-CNR), Dino Pedreschi (Univ. Pisa).

8. REFERENCES

[1] F. Angiulli and F. Fassetti. Very efficient mining of distance-based outliers. In *Procs of CIKM-2007*, pages 791–800, 2007.

[2] F. Angiulli, F. Fassetti, and L. Palopoli. Detecting outlying properties of exceptional objects. *ACM Trans. on Database Systems (TODS)*, page to appear, 2009.

[3] A. Arning, R. Agrawal, and P. Raghavan. A linear method for deviation detection in large databases. In *Procs of KDD-96*, pages 164–169, 1996.

[4] V. Barnett and T. Lewis. *Outliers in Statistical Data*. John Wiley & Sons, 1994.

[5] R. Bolton and D. Hand. Statistical fraud detection: A review. *Statistical Science*, 17(3):235–255, 2002.

[6] L. Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996.

[7] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander. Lof: Identifying density-based local outliers. In *Procs of SIGMOD-2000*, pages 93–104, 2000.

[8] N. Chawla et al. Smoteboost: improving prediction of the minority class in boosting. In *Procs. of PKDD-2003*, pages 107–119, 2003.

[9] N. Chawla, N. Japkowicz, and A. Kolcz. Special issue on learning from imbalanced datasets. *SIGKDD Explorations*, 6(1), 2004.

[10] N. Chawla and G. Karakoulas. Learning from labeled and unlabeled data: An empirical study across techniques and domains. *Journal of Artificial Intelligence Research*, 23:331–366, 2005.

[11] W. W. Cohen. Fast effective rule induction. In *Procs of ICML-95*, pages 115–123, 1995.

[12] P. Domingos. Metacost: A general method for making classifiers cost-sensitive. In *Procs. of KDD-99*, pages 155–164, 1999.

[13] R. Holte, L. Acker, and B. Porter. Concept learning and the problem of small disjuncts. In *Procs. 11th IJCAI Conf.*, pages 813–818, 1989.

[14] M. Keen and S. Smith. Vat fraud and evasion: What do we know, what can be done. Technical Report WP/07/31, International Monetary Fund, 2007.

[15] E. M. Knorr and R. T. Ng. Algorithms for mining distance-based outliers in large datasets. In *Procs of VLDB-98*, pages 392–403, 1998.

[16] W. S. Lee and B. Liu. Learning with positive and unlabeled examples using weighted logistic regression. In *Procs of ICML-2003*, pages 448–455, 2003.

[17] S. Papadimitriou, H. Kitagawa, P. B. Gibbons, and C. Faloutsos. Loci: Fast outlier detection using the local correlation integral. In *Procs of ICDE-2003*, pages 315–326, 2003.

[18] F. Provost and T. Fawcett. Robust classification for imprecise environments. *Machine Learning*, 42(3):203–231, 2001.

[19] J. R. Quinlan. *C4.5: programs for machine learning*. Morgan Kaufmann, 1993.

[20] R. E. Schapire. The strength of weak learnability. *Machine Learning*, 5:197–227, 1990.

[21] D. M. Tax. *One-Class Classification*. PhD thesis, Delft University of Technology, 2001.

[22] G. Weiss. Mining with rarity: a unifying framework. *SIGKDD Explorations*, 6(1):7–19, 2004.

[23] G. Weiss and H. Hirsh. A quantitative study of small disjuncts. In *Procs 16th AAAI Conf.*, pages 665–670, 2000.

[24] I. H. Witten and E. Frank. *Data Mining: Practical Machine Learning Tools*. Morgan Kaufmann, 2005.

Appendix

THEOREM 8.1. *The SBR problem is NP-hard.*

PROOF. Consider the decision version of the problem at hand. Let D be a dataset, $\mathcal{R}$ a set of rules and X a user defined threshold. The SBR_D problem is: *let k be a real number, find the subset $\mathcal{R}'$ of $\mathcal{R}$ such that $|\mathcal{R}'(D)| = X$ and $\gamma(\mathcal{R}'(D)) \geq k$.* Next, it is proved that the SBR_D problem is NP-complete.

The proof is given by reduction from the DOMINATING SET problem, which is well-known to be NP-complete. The dominating set problem is the following: *given a graph $G = \langle V, E\rangle$, with V a set of nodes and E a set of edges, and an integer h: is there a subset W of V such that $|W| \leq h$ and every node in $V \setminus W$ is joined to at least one node in W by an edge in E?*

Starting from a graph G, a dataset D^G can be built as follows. Let $A = \{A_1, \ldots, A_n\}$ be a set of n attributes, where n is equal to $|V|$, and let c be the class attribute. D^G is composed by $2n^2 + n$ tuples. In particular, we distinguish among three groups of tuples:

Group 1: for each node i in V there are in D^G n identical tuples $t^i_1, \ldots, t^i_n$ such that $t^i_j[c] = 0$, and $t^i_j[A_k] = 0 \forall k \neq i$, $t^i_j[A_i] = 1$;

Group 2: for each node i in V there are in D^G:

1. a tuple t'^i_1 such that $t'^i_j[c] = 1$, and $t'^i_1[A_k] = 0, \forall k \neq i$, $t'^i_1[A_i] = -1$
2. $n-1$ identical tuple $t'^i_2, \ldots, t'^i_n$ such that $t'^i_j[c] = 0$, and $t'^i_j[A_k] = 0, \forall k \neq i$, $t'^i_j[A_i] = -1$;

Group 3: there are in D^G $2n$ tuples $t^0_1, \ldots, t^0_n, t^0_{n+1}, \ldots, t^0_{2n}$, such that $t^0_j[c] = 1$, and $t^0_j[A_k] = t^0_{j+n}[A_k] = 1$ if $k = j$ or there exists an edges joining the node $1 \leq j \leq n$ and the node $1 \leq k \leq n$ in E, $t^0_j[A_k] = t^0_{j+n}[A_k] = 0$ otherwise.

Group 1 contain n^2 tuples, Group 2 contain n^2 tuples, and Group 3 contain n tuples.

An example of such a reduction is reported in Figure 7.

Consider now the set of rules $\mathcal{R}^G = \{r_1, \ldots, r_n, r_{n+1}, \ldots, r_{2n}\}$ consisting in $2n$ rules; specifically, for each $i \in [1, n]$ there are in $\mathcal{R}$ two rules $r_i : A_i \in \{1\} \rightarrow c = 1$, and $r_{i+n} : A_i \in \{-1\} \rightarrow c = 1$. The former (resp., the latter) n rules are called positive (resp., negative) rules. Moreover, given a node $v_i \in V$ we refer to the rule $r_i : A_i \in \{1\} \rightarrow c = 1$ as the positive rule associated with v_i; whereas we refer to the rule $r_{i+n} : A_i \in \{-1\} \rightarrow c = 1$ as the negative rule associated with v_i.

First of all, note that each negative rule r^- selects exactly n tuples, and no other rule in $\mathcal{R}$ can select these same tuples. Conversely, each positive rule r^+ selects n tuples of Group 1 and at least one tuple of Group 3. The tuples of Group 1 selected by r^+ cannot be selected by any other rule in $\mathcal{R}$, whereas the one or more tuples of Group 3 selected by r^+ can be selected by other rules.

Consider the example in Figure 7, and the rule $A_2 \in \{-1\} \rightarrow c = 1$. Such a rule selects five tuples: $t'^2_1, t'^2_2, t'^2_3, t'^2_4, t'^2_5$, that cannot be selected by any other tuple in $\mathcal{R}$. Consider now the rule $A_2 \in \{1\} \rightarrow c = 1$. Such a rule selects nine tuples: $t^2_1, t^2_2, t^2_3, t^2_4, t^2_5$ and $t^0_1, t^0_2, t^0_3, t^0_4$. Note that, while the tuples t^2_i cannot be selected by any other tuples, the tuple t^0_1 can be selected, for example, also by the rule $A_1 \in \{1\} \rightarrow c = 1$, and the tuples t^0_2, t^0_3, t^0_4 can be selected, for example, also by the rule $A_3 \in \{1\} \rightarrow c = 1$.

As for the true positives, each negative rule selects one true positive and $n - 1$ false positives; each positive rule selects as many true positives as are the rule of the group 3 selected.

Let h be a fixed integer. Next, it is proved that G has a dominating set D with $|D| \leq h$ if and only if there exists a solution for the SBR_D problem on the dataset G with the set of rules $\mathcal{R}$ and parameters $X = n^2 + n$ and $k = \frac{2n-h}{n^2+n}$.

As the first step, it is proved that if G has a dominating set D with $|D| \leq h$ then there exists a solution for the SBR_D problem.

Let $D = \{v_{i_1}, \ldots, v_{i_\ell}\}$ be a dominating set for G, with $\ell \leq h$. Consider the subset of rules $\mathcal{R}^D$ consisting in the positive rules associated with the nodes in D and the negative rules associated with the nodes in $V \setminus D$.

Consider Figure 7 again, and the dominating set $D = \{v_2, v_3\}$. Then, the set $\mathcal{R}^D$ is composed by the following rules:

r_6: $A_1 \in \{-1\} \rightarrow c = 1$,
r_2: $A_2 \in \{1\} \rightarrow c = 1$,
r_3: $A_3 \in \{1\} \rightarrow c = 1$,
r_4: $A_4 \in \{-1\} \rightarrow c = 1$, and
r_{10}: $A_5 \in \{-1\} \rightarrow c = 1$.

Since in $\mathcal{R}^D$ there are $n - \ell$ negative rules and ℓ positive rules, $\mathcal{R}^D$ selects at least $(n - \ell) \cdot n + \ell \cdot n = n^2$ rules. Moreover, since the nodes in D composed a dominating set, each tuple of the Group 3 has a 1 in at least one column associated with a node in D. Then, the positive rules of $\mathcal{R}^D$ select all the n tuples of the Group 3. Summarizing, $\mathcal{R}^D$ selects $n^2 + n$ rules, then the constraint on the parameter X is complied with.

As for the true positives selected by $\mathcal{R}^D$, the $n-\ell$ negative rules select $n - \ell$ true positives, whereas the ℓ positive rules selects n true positives (all the tuples of Group 3). Hence, the global confidence of $\mathcal{R}^D$ is $\frac{2n-\ell}{n^2+n}$, then the constraint on the global confidence is complied with, and this concludes the first step of the proof.

As a second step, it is proved that if there exists a solution $\mathcal{R}^*$ for the SBR_D problem with the set of rules $\mathcal{R}^G$ such that $\mathcal{R}^*(D^G) = n^2 + n$ and $\gamma(\mathcal{R}^*) \geq \frac{2n-h}{n^2+n}$, then G has a dominating set D such that $|D| \leq h$.

First, it is proved that $\mathcal{R}^*$ is composed by exactly n rules. Suppose that in $\mathcal{R}^*$ there are $n - 1$ rules. Such rules can select at most all the n tuples of Group 3, and $(n - 1) \cdot n$ tuples of Group 1 or 2. Then globally the rules in $\mathcal{R}^*$ select $(n-1)\cdot n+n = n^2$ tuples which is lower than X. Conversely, suppose that in $\mathcal{R}^*$ there are $n + 1$ tuples. Such rules at least are n negative rules plus one positive rules. Since each positive rule select at least $n + 1$ tuples, globally, the rules in $\mathcal{R}^*$ select $n2 + (n + 1)$ tuples which is larger than X.

Moreover, in order for $\mathcal{R}^*$ to select $n^2 + n$ tuples, some positive rules must be in $\mathcal{R}^*$. Indeed, if $\mathcal{R}^*$ were composed of only negative rules, it would select at most n^2 tuples.

Let m be the number of positive rules in $\mathcal{R}^*$. Then, in $\mathcal{R}^*$ there are $n - m$ negative rules. Hence, $\mathcal{R}^*$ select $[(n - m) \cdot n] + [m \cdot n + \nu]$ tuples, where the first term is due to the negative rules, while the second term is due to the positive rules, and ν is the number of tuples of Group 3 selected by the rules in $\mathcal{R}'$.

Since the number of tuples selected by $\mathcal{R}^*$ must be equal to X, $[(n-m)\cdot n]+[m\cdot n+\nu]$ must be equal to n^2+n, and then ν must be equal to n. This implies that the positive rules in $\mathcal{R}^*$ select all the tuples of Group 3.

Summarizing, in $\mathcal{R}^*$ there are m positive rules and $n-m$ negative rules, and the m positive rules are able to select all the tuples of Group 3.

As for the true positives selected by $\mathcal{R}^*$, it holds that the each negative rule of $\mathcal{R}^*$ selects a true positive, then globally the negative rules select $n-m$ true positives; moreover, the positive rules select as many true positive as tuples of Group 3 are selected, then they select n true positives. Therefore, the confidence of $\mathcal{R}^*$ is $\frac{n+n-m}{n^2+n}$. Since by hypothesis $\gamma(\mathcal{R}^*) \geq \frac{2n-h}{n^+n}$, it follows that $m \leq h$.

Summarizing, we prove that if there exists a solution $\mathcal{R}^*$ for the SBR_D problem with the set of rules $\mathcal{R}^G$ such that $\mathcal{R}^*(D^G) = n^2+n$ and $\gamma(\mathcal{R}^*) \geq \frac{2n-h}{n^2+n}$, then the number of positive rules in $\mathcal{R}^G$ is lower than h.

Next, to conclude the proof, we obtain a dominating set from $\mathcal{R}^*$. Consider the positive rules in $\mathcal{R}^*$. It follows form the above discussion that they are m with $m \leq h$ and that they select all the tuples of Group 3.

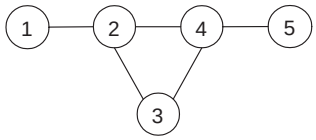
Hence, consider the set D of m nodes v_i such that $A_i \in \{1\} \rightarrow c = 1$ is a positive rule in $\mathcal{R}^*$.

Since each tuple of Group 3 is selected, it follows that:

$$\forall j \in [1,n] \exists i \mid t_j^0[A_i] = 1 \text{ and } A_i \in \{1\} \rightarrow c = 1 \in \mathcal{R}^*.$$

It follows, by construction, that the nodes in D are joined to all the nodes of the graph by at least one edge. Hence, D is a dominating set and $|D|$ is equal to $m \leq h$.

As an immediate consequence of what above proved, it follows that the SBR problem is NP-hard. □



ID	A_1	A_2	A_3	A_4	A_5	c
t_1^1	1	0	0	0	0	0
t_1^1	1	0	0	0	0	0
t_2^1	1	0	0	0	0	0
t_3^1	1	0	0	0	0	0
t_4^1	1	0	0	0	0	0
t_5^1	1	0	0	0	0	0
t_1^2	0	1	0	0	0	0
t_1^2	0	1	0	0	0	0
t_2^2	0	1	0	0	0	0
t_3^2	0	1	0	0	0	0
t_4^2	0	1	0	0	0	0
t_5^2	0	1	0	0	0	0
t_1^3	0	0	1	0	0	0
t_1^3	0	0	1	0	0	0
t_2^3	0	0	1	0	0	0
t_3^3	0	0	1	0	0	0
t_4^3	0	0	1	0	0	0
t_5^3	0	0	1	0	0	0
t_1^4	0	0	0	1	0	0
t_1^4	0	0	0	1	0	0
t_2^4	0	0	0	1	0	0
t_3^4	0	0	0	1	0	0
t_4^4	0	0	0	1	0	0
t_5^4	0	0	0	1	0	0
t_1^5	0	0	0	0	1	0
t_1^5	0	0	0	0	1	0
t_2^5	0	0	0	0	1	0
t_3^5	0	0	0	0	1	0
t_4^5	0	0	0	0	1	0
t_5^5	0	0	0	0	1	0
t'^1_1	-1	0	0	0	0	1
t'^1_2	-1	0	0	0	0	0
t'^1_3	-1	0	0	0	0	0
t'^1_4	-1	0	0	0	0	0
t'^1_5	-1	0	0	0	0	0
t'^2_1	0	-1	0	0	0	1
t'^2_2	0	-1	0	0	0	0
t'^2_3	0	-1	0	0	0	0
t'^2_4	0	-1	0	0	0	0
t'^2_5	0	-1	0	0	0	0
t'^3_1	0	0	-1	0	0	1
t'^3_2	0	0	-1	0	0	0
t'^3_3	0	0	-1	0	0	0
t'^3_4	0	0	-1	0	0	0
t'^3_5	0	0	-1	0	0	0
t'^4_1	0	0	0	-1	0	1
t'^4_2	0	0	0	-1	0	0
t'^4_3	0	0	0	-1	0	0
t'^4_4	0	0	0	-1	0	0
t'^4_5	0	0	0	-1	0	0
t'^5_1	0	0	0	0	-1	1
t'^5_2	0	0	0	0	-1	0
t'^5_3	0	0	0	0	-1	0
t'^5_4	0	0	0	0	-1	0
t'^5_5	0	0	0	0	-1	0
t_1^0	1	1	0	0	0	1
t_1^0	1	1	1	1	0	1
t_2^0	0	1	1	1	1	1
t_3^0	0	1	1	1	0	1
t_4^0	0	0	1	0	1	1

7: Example of the reduction employed in Theorem 8.1.