High Quality True-Positive Prediction for Fiscal Fraud Detection

S. Basta, F. Fassetti, F. Giannotti, M. Guarascio. G. Manco
G. Papi, D. Pedreschi, S. Pisani, L. Spinsanti

Outline

- Scenario and Motivation
- DIVA Overview
  - Solution Proposed
  - Scoring Criteria
  - Multi-purpose objectives
- Sniper Core
  - Generating Rule
  - Merging Rule
- Evaluation
- Conclusion

The Context: VAT frauds in Italy

- DIVA - A joint initiative involving academic researchers, experts on fiscal laws, IT Professionals
- Main objective:
  - To tackle the VAT Fraud Detection issue raised by the credit mechanism via the adoption of data mining techniques.

Scenario

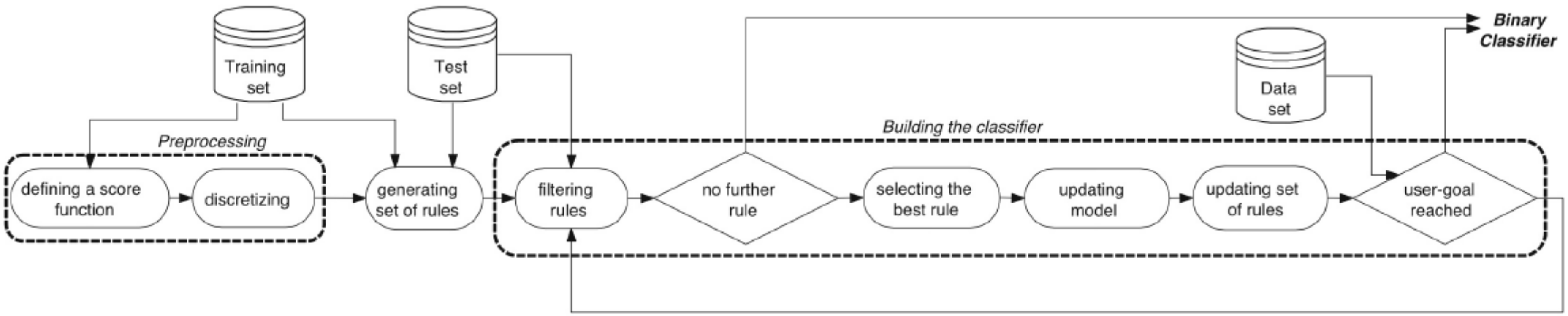
- Several challenges, both from a scientific and a practical point of view:
  - Sample selection bias
    - Audited subjects are not randomly chosen
    - Highly skewed data
      - Positive subjects larger than non-defrauders in audit data
  - Imprecise settings
    - Inaccurate, incomplete, and irrelevant data attributes
  - Only 0.004% of population audited

Motivation

- Classical approaches to the problem of fraud detection are not very effective:
  - Rule-Based classifiers are preferable for interpretability, but
    - Poor predictive accuracy in highly imprecise learning settings
    - Class-imbalance problem
  - Cost-sensitive classification and meta-learning approaches suffer from low interpretability

The proposal: Sniper as a meta-learner

- The core of the Sniper technique is the extraction of a binary rule-based classifier able to identify X topmost defrauders
  - Based on the combined use of local models and the definition of multi-objective functions.



DIVA Overview

- The data made available by the agency consisted of about 34 million VAT declarations spread over 5 years.
- Data contain general 'demographic' information, plus specific information about VAT declarations.
- As a result of a data understanding process conducted jointly with domain experts, we chose a total of 135 such features and 45,442 audited subjects.

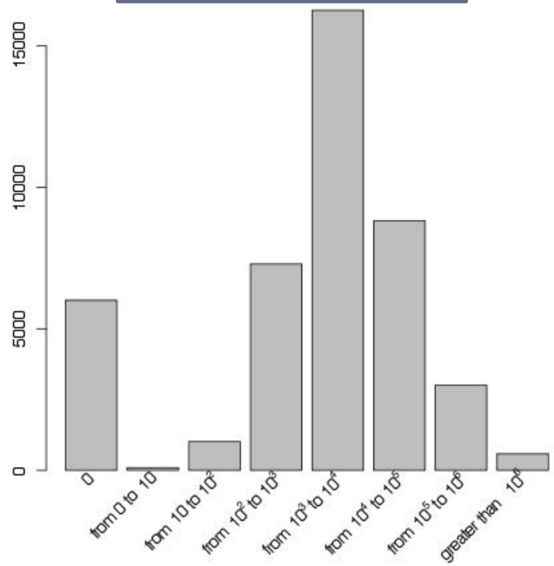
Scoring individuals

- A multi-purpose modeling strategy, aiming at characterizing the exceptionalness and interestingness of an individual
  - PROFITABILITY: The amount of VAT fraud
    - The higher, the better
  - EQUITY
    - Low amounts do not necessarily correspond to meaningless fraudsters. The amount of fraud is relevant related to their business volume (1.000eur on 10.000eur is better than 1.000eur on 100.000eur)
  - EFFICIENCY
    - Scoring and detection should be sensitive to total/partial frauds (underclaring 200eur declaring 2.000eur is less dignificant than underclaring 200eur declaring 200eur)

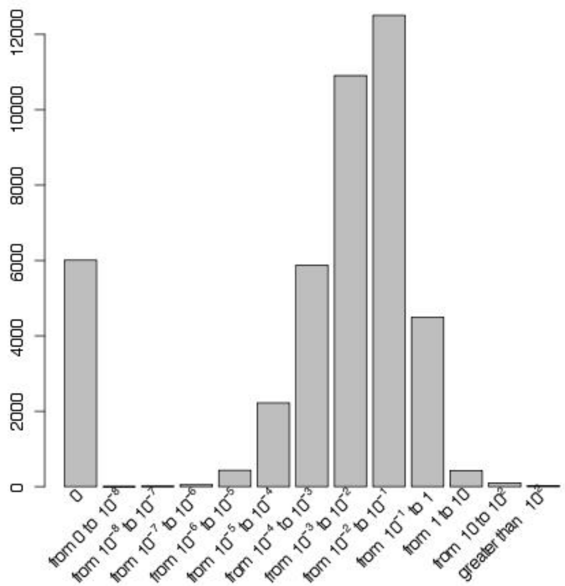
Issues

- Need to face a trade-off among profitability, equity and efficiency
  - Solution: a combination of baseline functions
  - AND,OR, FUZZY_AND, FUZZY_OR

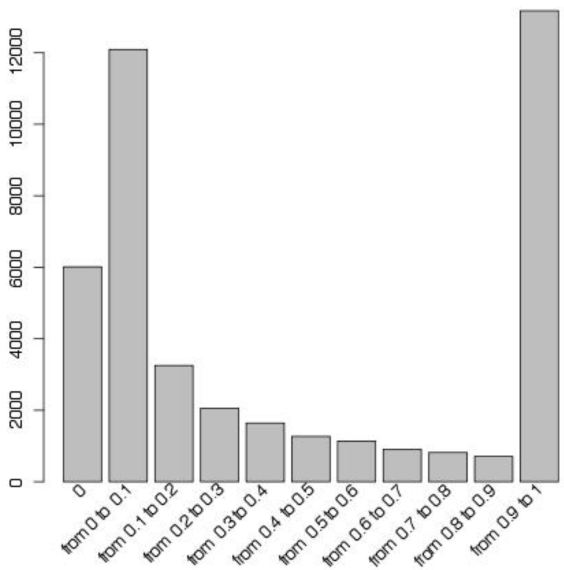
profitability



efficiency



equity



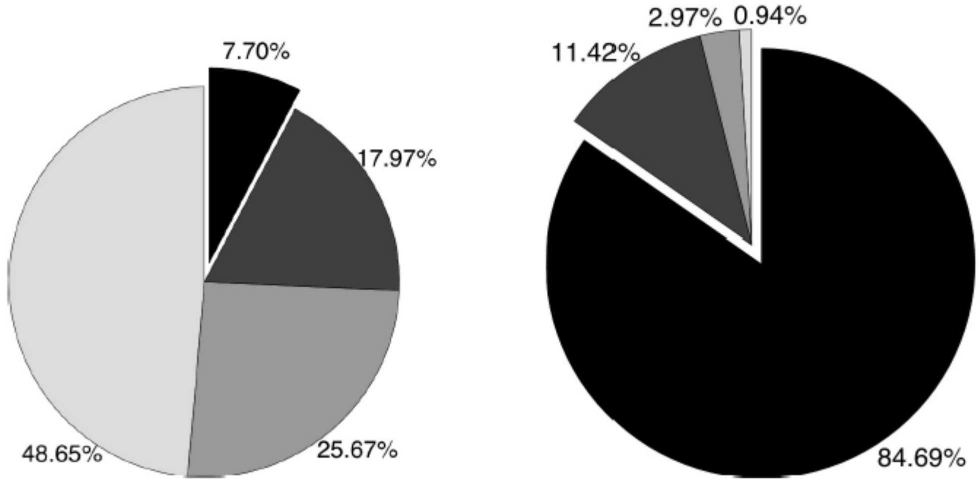
The Fuzzy combination

- Two different objective functions, four main classes

harmonization f. weight

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

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



Subject partitioning Retrieved fraud

Score function results

Generating rules

- Sniper builds a hybrid classifier, resulting from the combination of the whole set of classifiers trained over the training set
- Advantages:
  - Separate model construction from model selection
  - Model construction
    - Several different strategies are attempted to build models focused on local peculiarities of the top class
  - Model selection
    - Several local fragments can be selected or discarded if the global accuracy improves

Merging Rules

- A candidate ruleset R is obtained by merging all the rules returned by h classifiers modeling the top class

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

- 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 R that activates it.
- The model is distilled from R *by selecting accurate rules, and removing* inaccurate rules from R in a principled (confidence-based) way

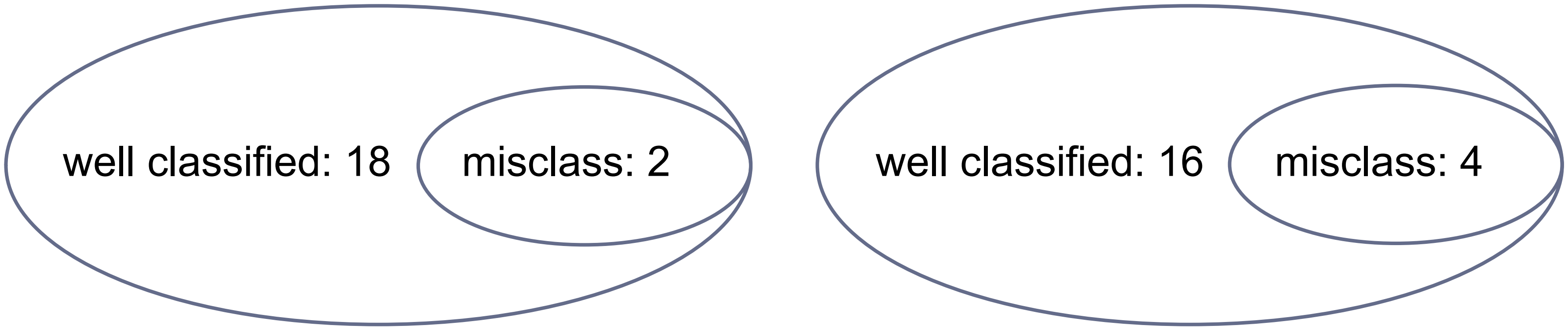
Building Ruleset

- Why we cannot just collect all the "good" rules from our classifiers?

$conf_{min} = 0.8$

Rule 1: sup=20 conf=0.9

Rule 2: sup=20 conf=0.8



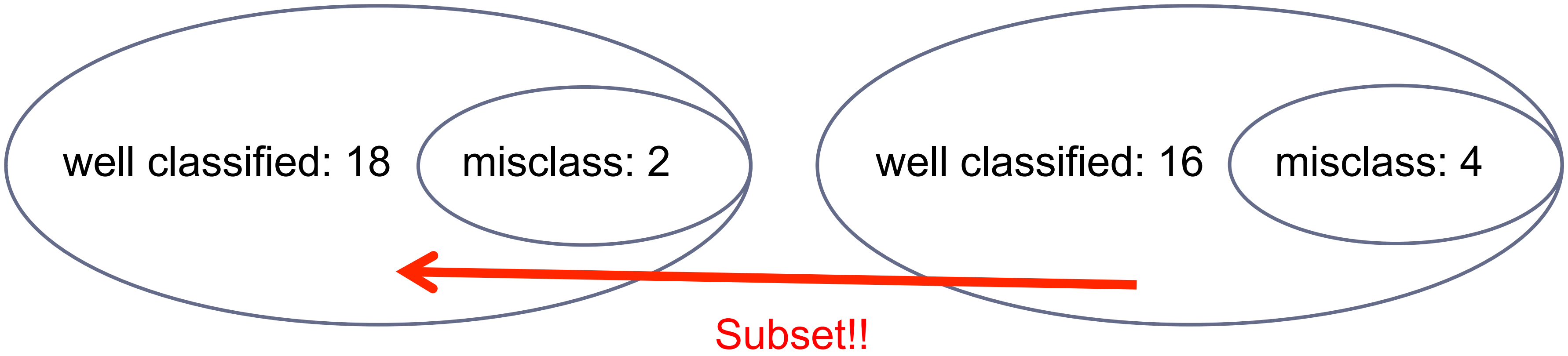
Building Ruleset

- Why we cannot just collect all the "good" rules from our classifiers?

$conf_{min} = 0.8$

Rule 1: sup=20 conf=0.9

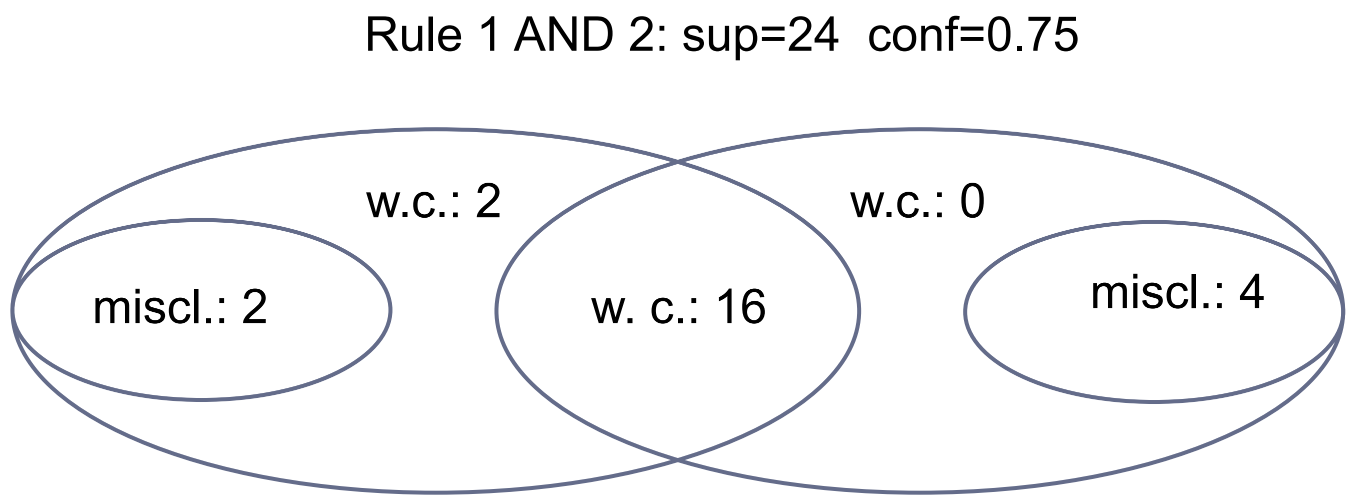
Rule 2: sup=20 conf=0.8



Building Ruleset

- Why we cannot just collect all the "good" rules from our classifiers?

$conf_{min} = 0.8$



Merging Rules

Input: A set of non-exclusive positive rules $\mathcal{R}$,
a confidence threshold $\gamma_{\min}$,
an integer X

Output: A model $\mathcal{M}$

Method:

$\mathcal{M} := \emptyset$

$\mathcal{R} := \left\{ r \in \mathcal{R} \mid \gamma(r) \geq \gamma_{\min} \right\}$

while $\mathcal{R} \neq \emptyset$ **do** *//first stop condition*

  $r^* := \arg\max_{r \in \mathcal{R}} \left\{ \gamma(r) \right\}$ *//select the best rule*

  $\mathcal{M} := \mathcal{M} \cup \{r^*\}$ *//update the current model*

  if $\mathcal{M}(D) \geq X$ **then** *//second stop condition*

    return $\mathcal{M}$

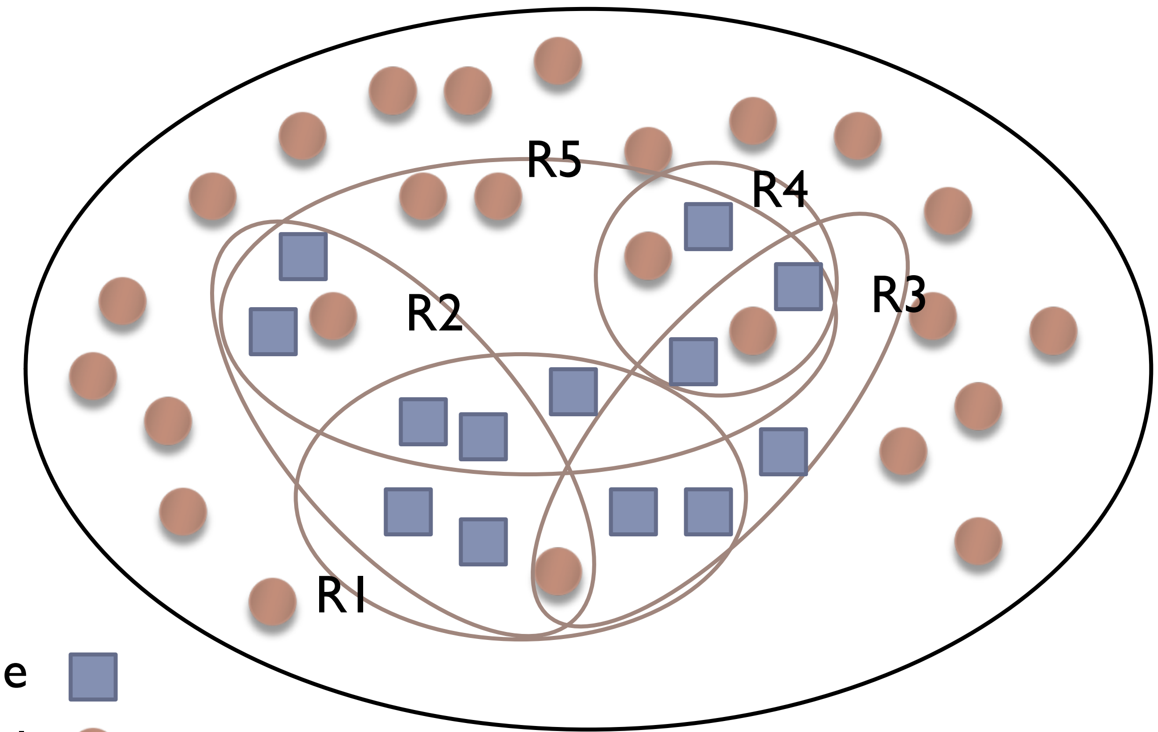
  $\mathcal{R}$ is updated by removing r^* and by replacing each rule r other than r^* with the rule r' if $\gamma(r') = \gamma_{\min}$, otherwise r is just removed from $\mathcal{R}$

return $\mathcal{M}$

Merging Rules: Example

- Assume $\gamma_{\min} = 60\%$
- Initially, $R = \{R1, R2, R3, R4, R5\}, M = \{\}$

Rule_ID	Confidence
R1	87,50%
R2	75%
R3	71,4%
R4	60%
R5	58,30%

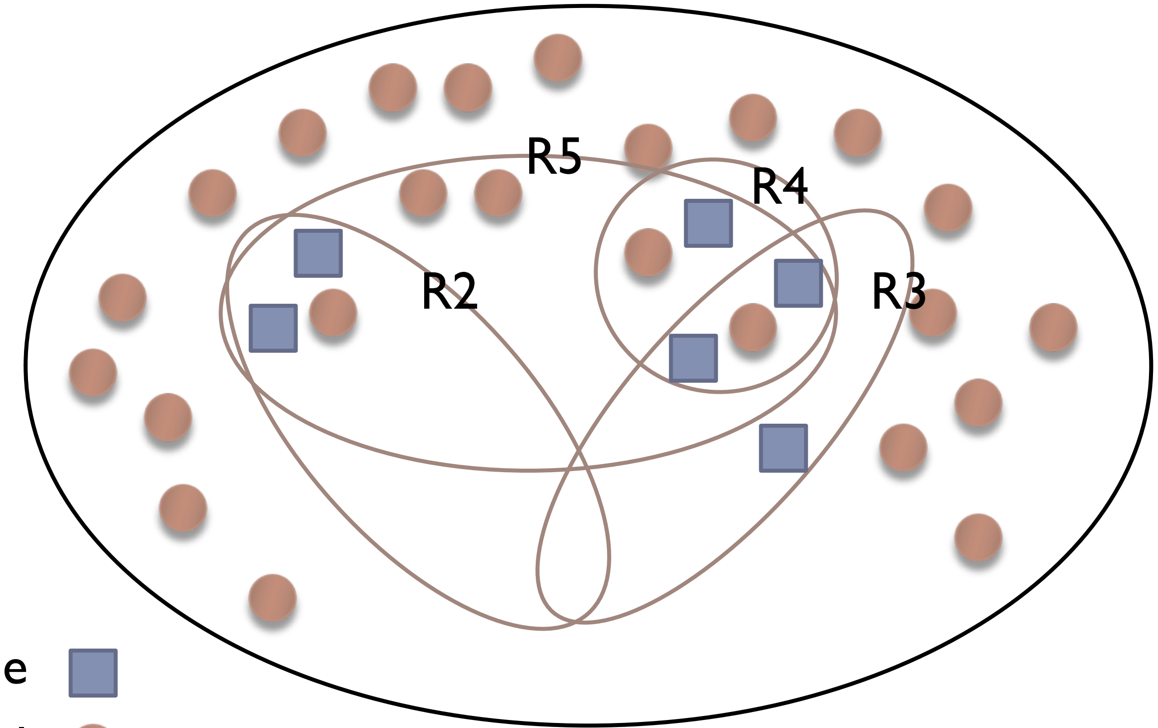


- Positive Example
- Negative Example

Merging Rules: Example

- $R = \{R2,R3,R4,R5\}$, $M=\{R1\}$

Rule_ID	Confidence
R2	66,6%
R3	75%
R4	60%
R5	50%

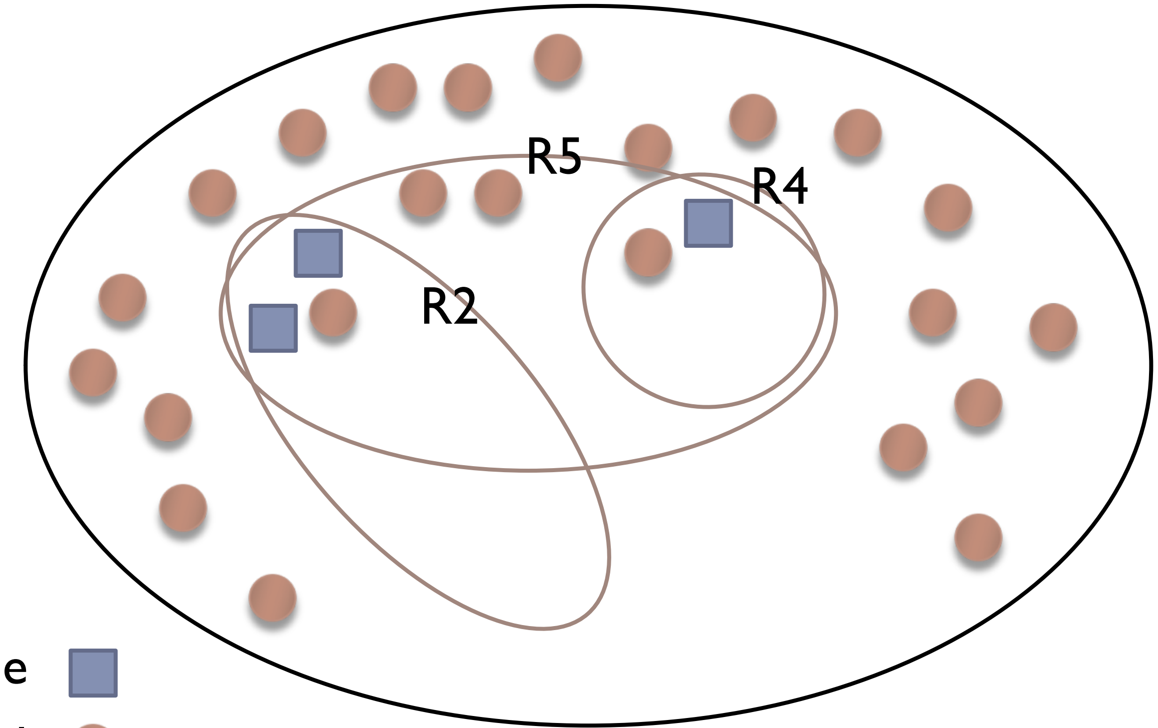


- Positive Example
- Negative Example

Merging Rules: Example

- $R = \{R2,R4,R5\}$, $M=\{R1,R3\}$

Rule_ID	Confidence
R2	66,6%
R4	50%
R5	42,8%

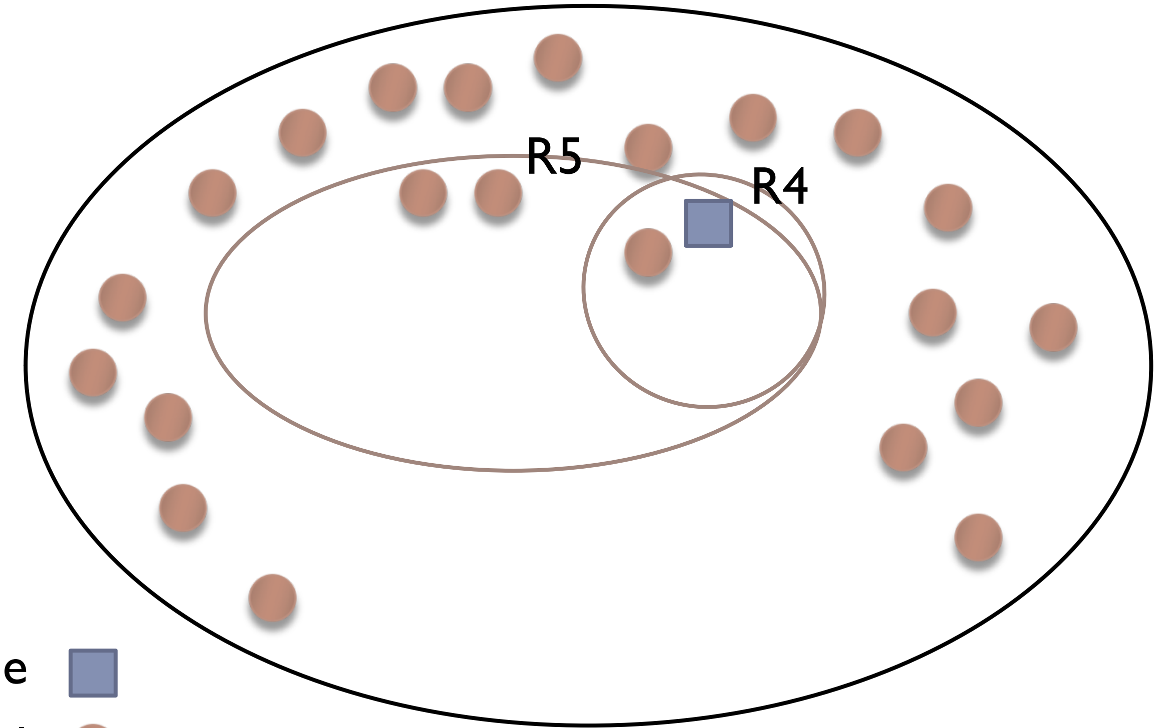


- Positive Example
- Negative Example

Merging Rules: Example

- $R = \{R4, R5\}, M = \{R1, R3, R2\}$

Rule_ID	Confidence
R4	50%
R5	25%



- Positive Example
- Negative Example

Evaluation

- We compared the results obtained from a single classifier against those obtained by Sniper in terms of confidence and support of the rules generated

classifier	*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
$\boldsymbol{S}$	8.78	80.41	9,840

(Partial) Results

- 1475 subjects identified
  - 276 subjects audited (feb-2010)
    - 147 in class 3 (53,26%)

- Mean Values:
  - Proficiency: 77.514,14
  - Equity: 32,5738
  - Efficiency: 0,4252