Data Minin and Fraud detection

DM2 – Scenario Fraud detection

Material integrated by slides of Chris Clifton

Pardue University

Problem definition: What is Fraud Detection?

- Identify wrongful actions
  - Is right and wrong universal?
  - If so, why not just prevent wrong actions
- Identify actions by the wrong people
- Identify *suspect* actions
  - Legal
  - But probably not right

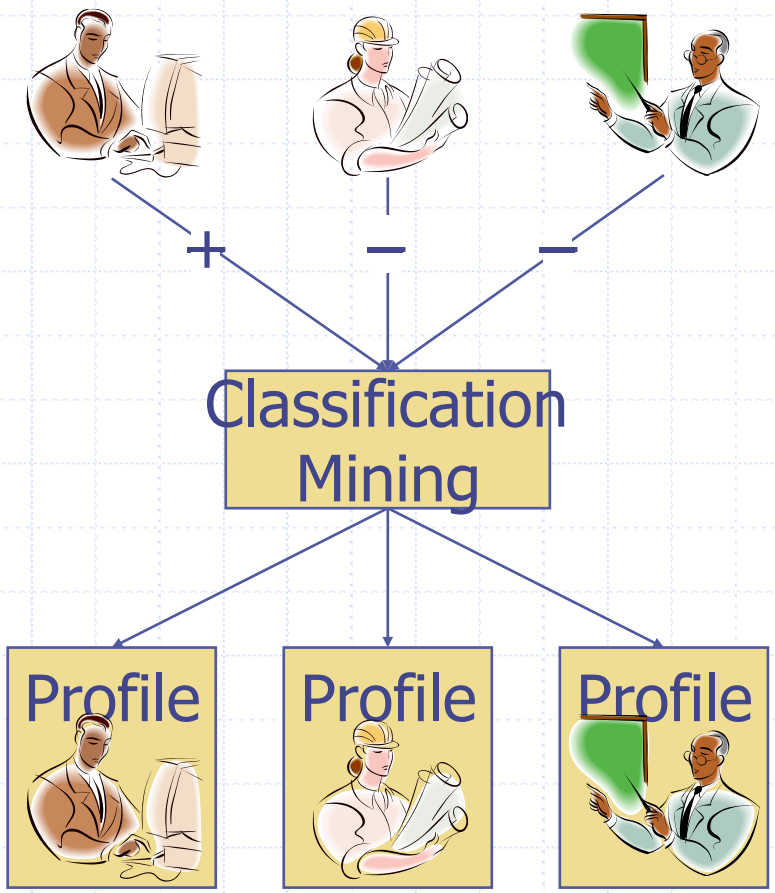
In Data Mining terms…

- Classification?
  - Classify into fraudulent and non-fraudulent behavior
  - What do we need to do this?
- Outlier Detection
  - Assume non-fraudulent behavior is normal
  - Find the exceptions
- Problems?

Solution: Differential Profiling

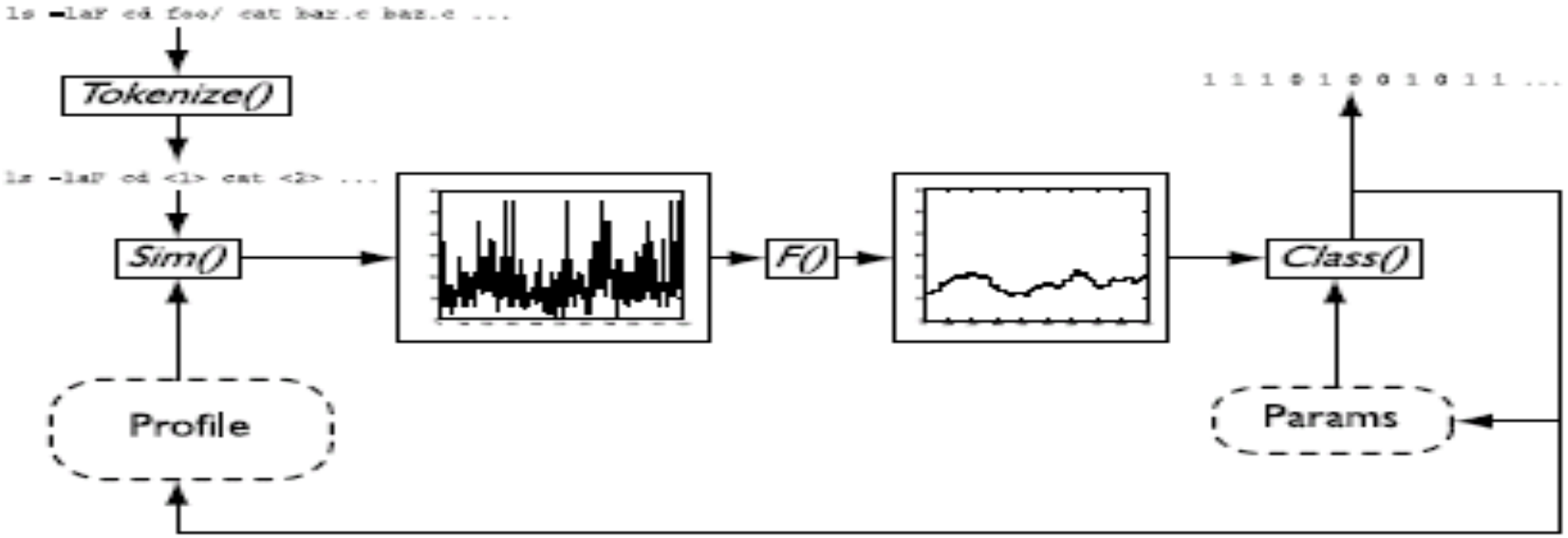
Solution: Differential Profiling

- Determine individual behavior
  - What is normal for the individual
  - What separates one individual from another
- Gives profile of individual behavior
- How do we do this?



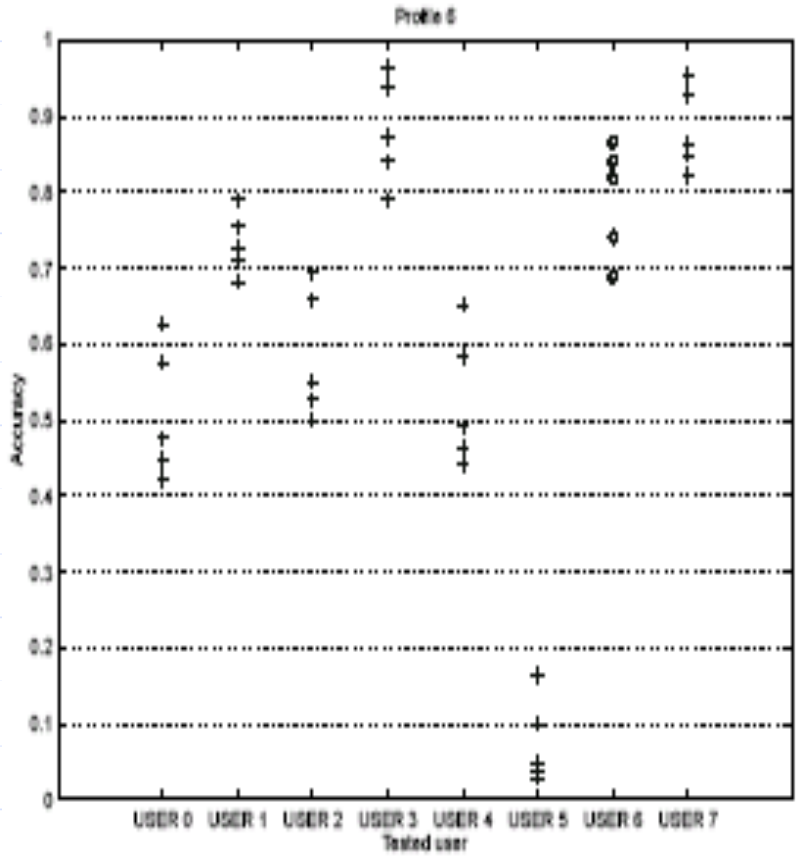
Has this been done? Intrusion Detection *(Lane&Brodley)*

- Profiled computer users based on command sequences
  - Command
  - Some (but not all) argument information
  - Sequence information

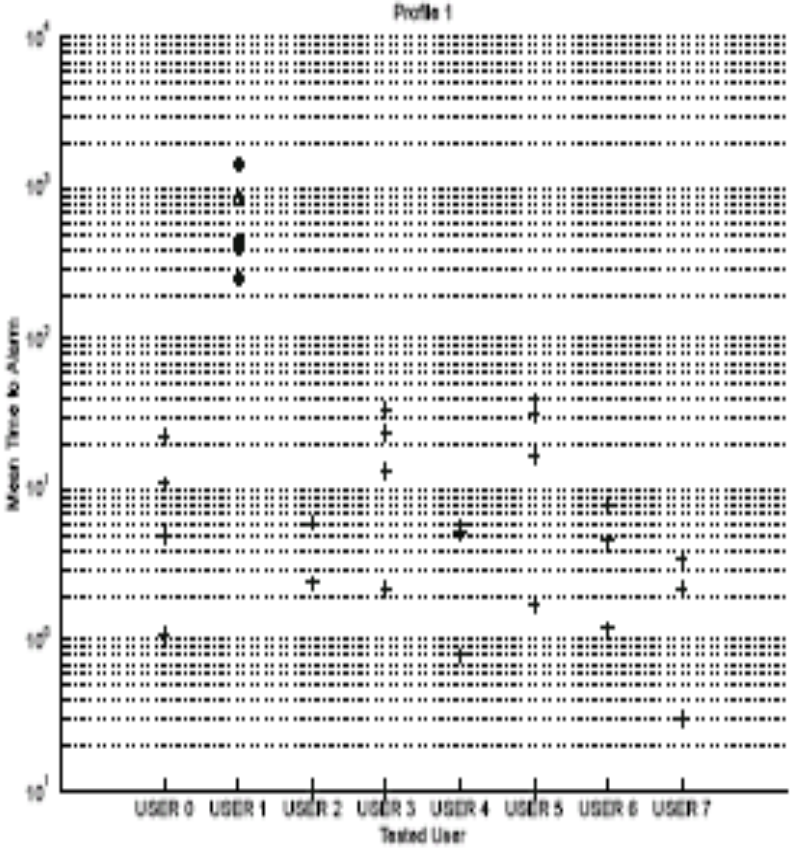


Results

Accuracy



Time to Alarm



Scaling Issues

- What happens with millions of users?
  - Credit card
  - Cell phone
- What about new users?
- Ideas?

Multi-user profiles

- Cluster users
- Develop profiles for clusters
  - E.g., differential profiling
- Old customers: Do they match profile for their cluster?
  - Allows wider range of acceptable behavior
- New customer: Do they match *any* profile?

Matching known fraud/non-compliance

- Which new cases are similar to known cases?
- How can we define similarity?
- How can we *rate* or *score* similarity?

Anomalies and irregularities

- How can we detect anomalous or unusual behavior?
- What do we mean by usual?
- Can we rate or score cases on their degree of anomaly?

Techniques used to identify fraud

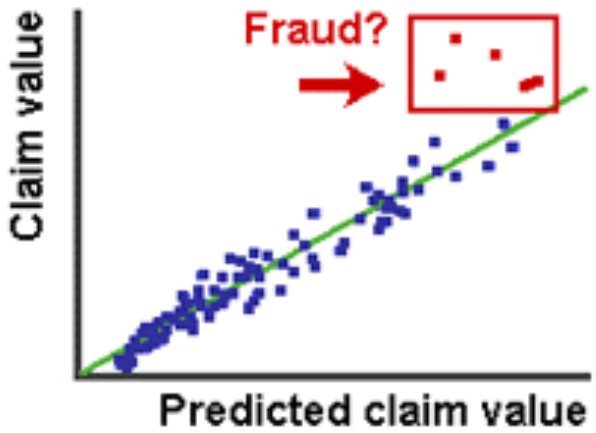
Predict and Classify

- Regression algorithms (predict numeric outcome): neural networks, CART, Regression, GLM
- Classification algorithms (predict symbolic outcome): CART, C5.0, logistic regression

Group and Find Associations

- Clustering/Grouping algorithms: K-means, Kohonen, 2Step, Factor analysis
- Association algorithms: apriori, GRI, Capri, Sequence

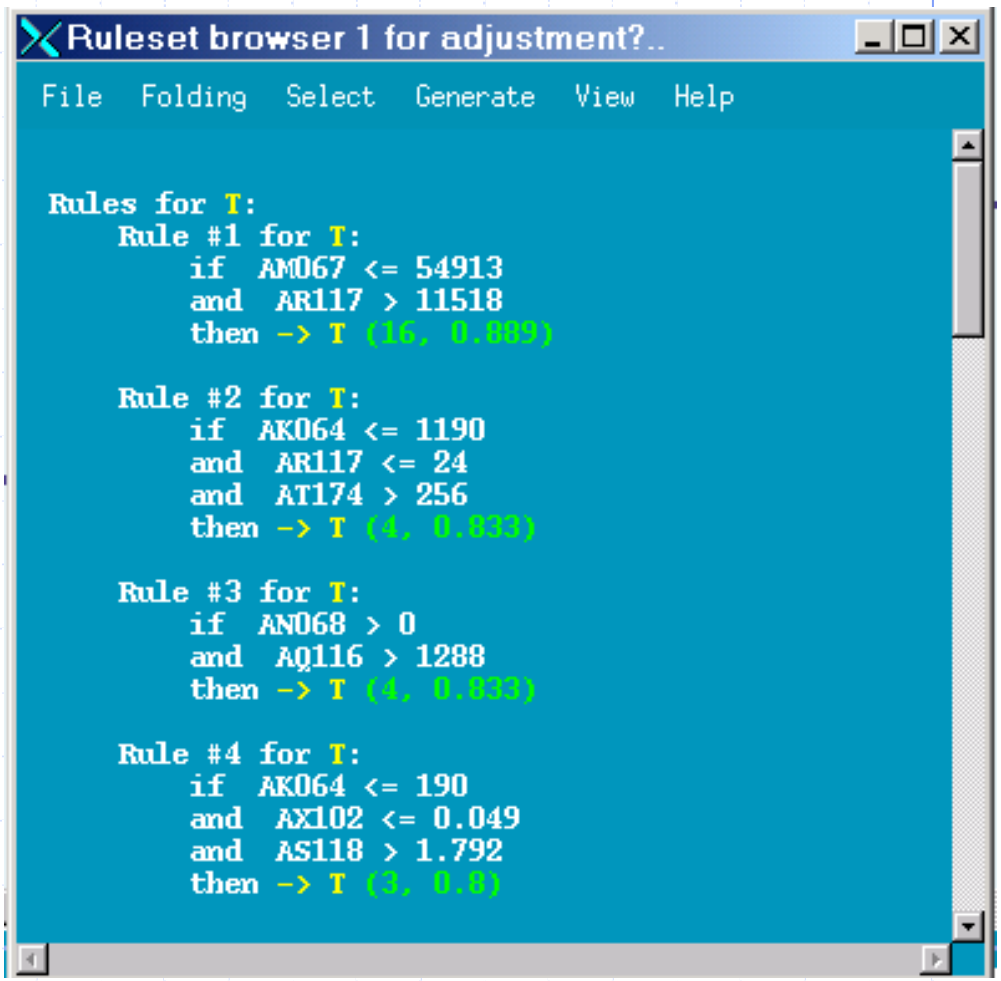
Techniques for finding fraud:



- Predict the expected value for a claim, compare that with the actual value of the claim.
- Those cases that fall far outside the expected range should be evaluated more closely

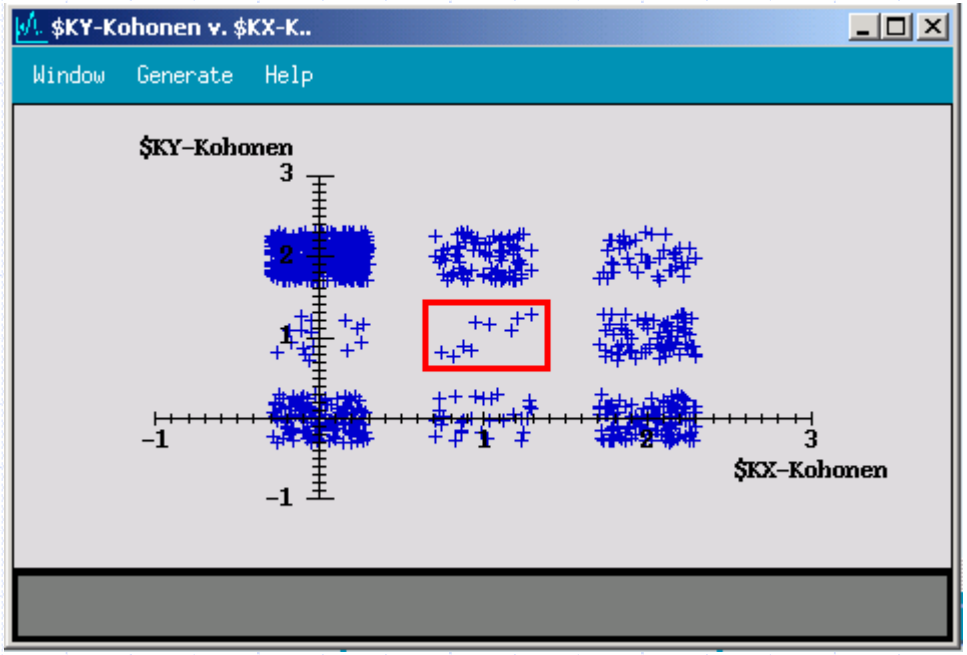
Techniques for finding fraud:

- Build a profile of the characteristics of fraudulent behavior.
- Pull out the cases that meet the historical characteristics of fraud.



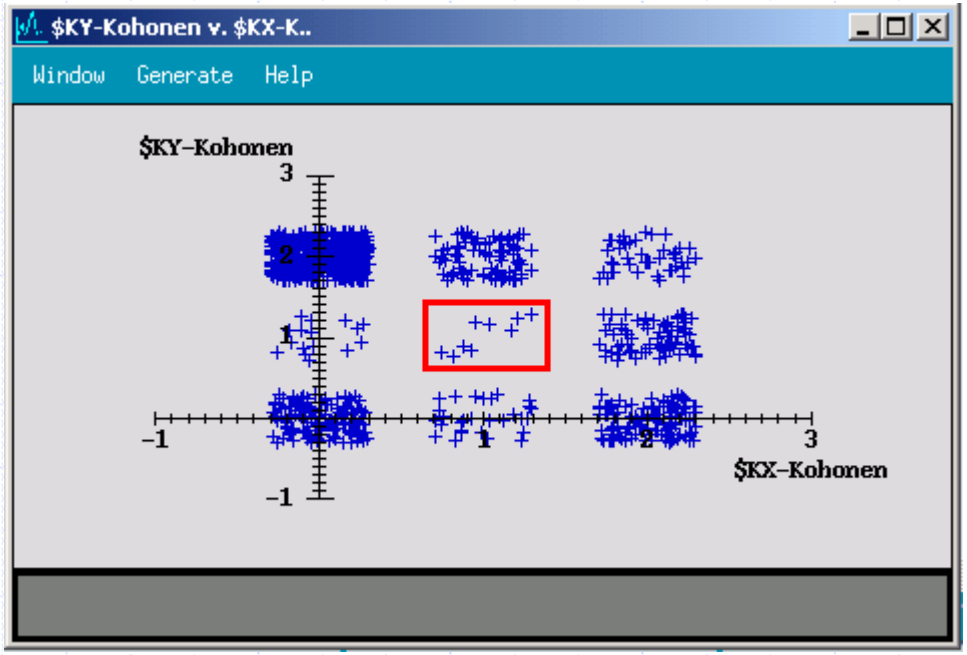
Techniques for finding fraud:

- ***Clustering and Associations***
  - Group behavior using a clustering algorithm
  - Find groups of events using the association algorithms
  - Identify outliers and investigate



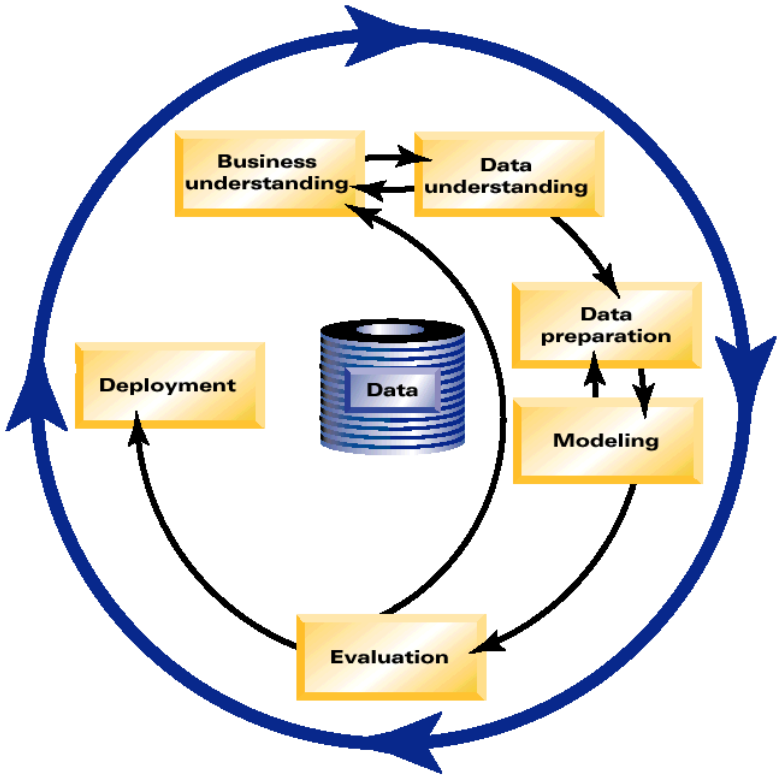
Techniques for finding fraud:

- ***Clustering and Associations***
  - Group behavior using a clustering algorithm
  - Find groups of events using the association algorithms
  - Identify outliers and investigate



Fraud detection using CRISP-DM

- Provides a systematic way to detect fraud and abuse
- Ensures auditing and investigative efforts are maximized
- Continually assesses and updat models to identify new emergir fraud patterns
- Leads to higher recoupments



Rilevamento di frodi fiscali e pianificazione degli accertamenti

Sorgente: Ministero delle Finanze

Progetto Sogei, KDD Lab. Pisa

Lotta all'evasione – Min. Finanze/SOGEI ('98-'99)

- **Pianificazione di accertamenti fiscali**
- **Obiettivo**: costruire un modello predittivo che individui una porzione di contribuenti su cui risulti vantaggioso effettuare un controllo fiscale.
  - Estrazione di **alberi di decisione**
- **Dataset**:
  - dati storici provenienti da fonti diverse (mod. 760, mod. 770, INPS, ENEL, SIP, Camere del Commercio)
  - dati storici sui risultati degli accertamenti pregressi.
- Variabile da predire: imposta recuperata al netto delle spese di accertamento.
- Valutazione dei modelli estratti rispetto ad **indici** generali (accuratezza) e specifici di dominio (redditività)

Rilevamento di frodi

- Obiettivo generale:
  - Determinare *modelli* per la previsione del comportamento fraudolento per:
  - Prevenire frodi future (rilevamento di frodi *on-line*)
  - Scoprire frodi passate (rilevamento frodi *a posteriori*)
- Obiettivo specifico:
  - Analizzare i dati storici sulle verifiche per pianificare verifiche future più EFFICACI

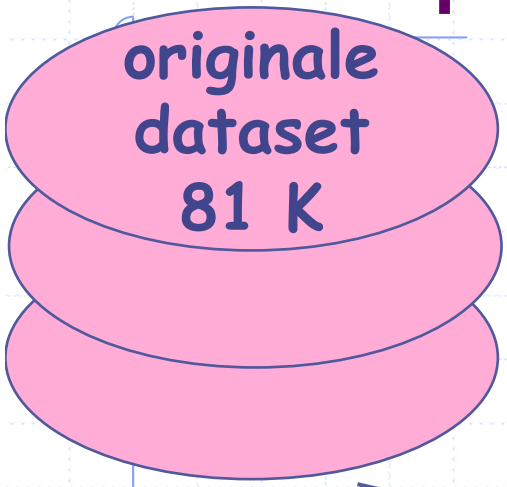
Pianificazione di verifiche

- C’ è un trade-off tra:
  - *Massimizzare i benefici della verifica*: selezionare quei contribuenti che massimizzano il recupero di tasse evase.
  - *Minimizzare il costo della verifica* : selezionare quei contribuenti che minimizzano le risorse necessarie alla verifica.

Available data sources

- Dataset: Dichiarazioni dei redditi, su una classe selezionata di aziende italiane integrate con altre sorgenti:
- Contributi INPS per dipendenti, consumi ENEL e telefonici..
- Dimensione: 80 K tuple, 175 numerici attribute.
- Un sottoinsieme di 4 K tuples corrisponde ad aziende verificate:
  - I risultati delle verifiche sono memorizzati nell'attributo: *recovery* (= *amount of evaded tax ascertained*)

Data preparation

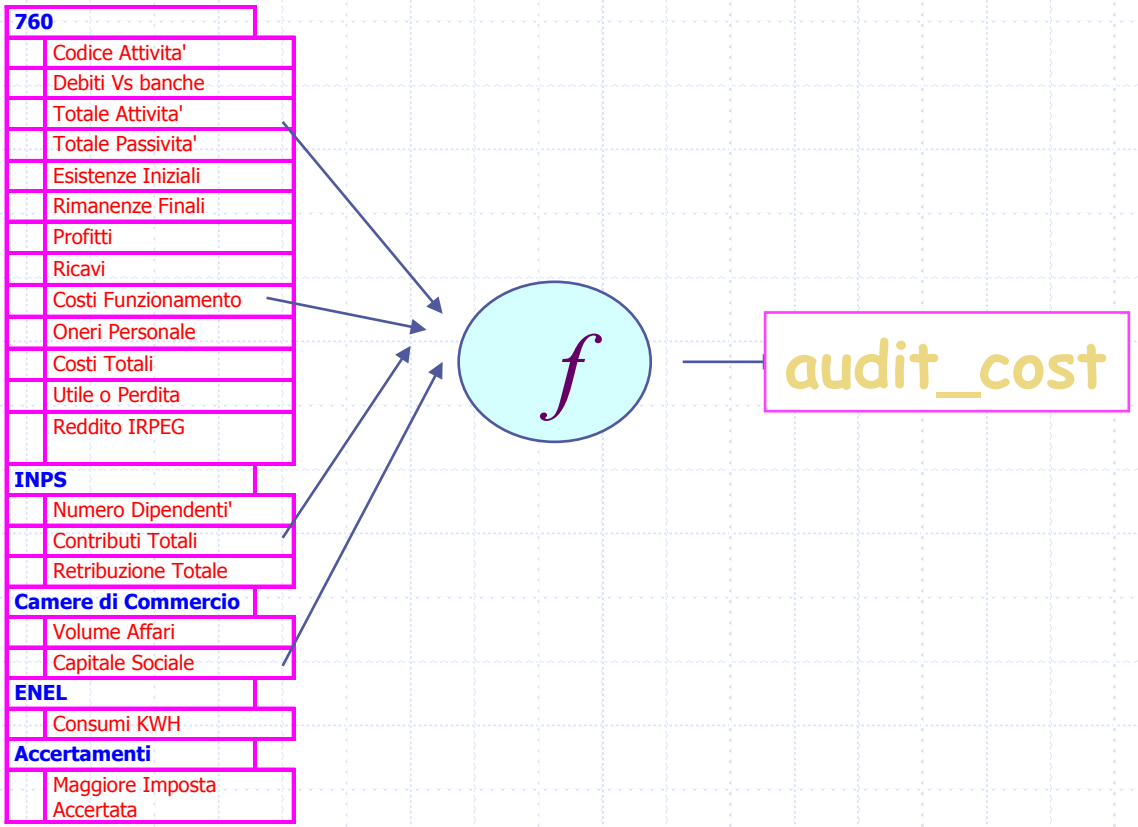


data consolidation
data cleaning
attribute selection

TAX DECLARATION	
	Codice Attivita'
	Debiti Vs banche
	Totale Attivita'
	Totale Passivita'
	Esistenze Iniziali
	Rimanenze Finali
	Profitti
	Ricavi
	Costi Funzionamento
	Oneri Personale
	Costi Totali
	Utile o Perdita
	Reddito IRPEG
SOCIAL BENEFITS	
	Numero Dipendenti'
	Contributi Totali
	Retribuzione Totale
OFFICIAL BUDGET	
	Volume Affari
	Capitale Sociale
ELECTRICITY BILLS	
	Consumi KWH
AUDIT	
	Recovery

Modello di costo

- si definisce l'indicatore audit_cost come funzione di altri attributi



Modello dei costi e variabile target

- Recupero di una verifica
  - actual_recovery = recovery - audit_cost
- La variabile target (class label) della nostra analisi: Class of Actual Recovery (c.a.r.):
- *c.a.r. =*

$$\begin{cases} \textit{negative} & \textbf{\textit{if}} \quad \textit{actual\_recovery} \leq 0 \\ \textit{positive} & \textbf{\textit{if}} \quad \textit{actual\_recovery} > 0. \end{cases}$$

Indicatori di qualità

- Si costruiscono vari classificatori che sono valutati secondo diverse metriche:
- Domain-independent indicators
  - confusion matrix
  - misclassification rate
- Domain-dependent indicators
  - audit #
  - actual recovery
  - profitability
  - relevance

Indicatori Domain-dependent

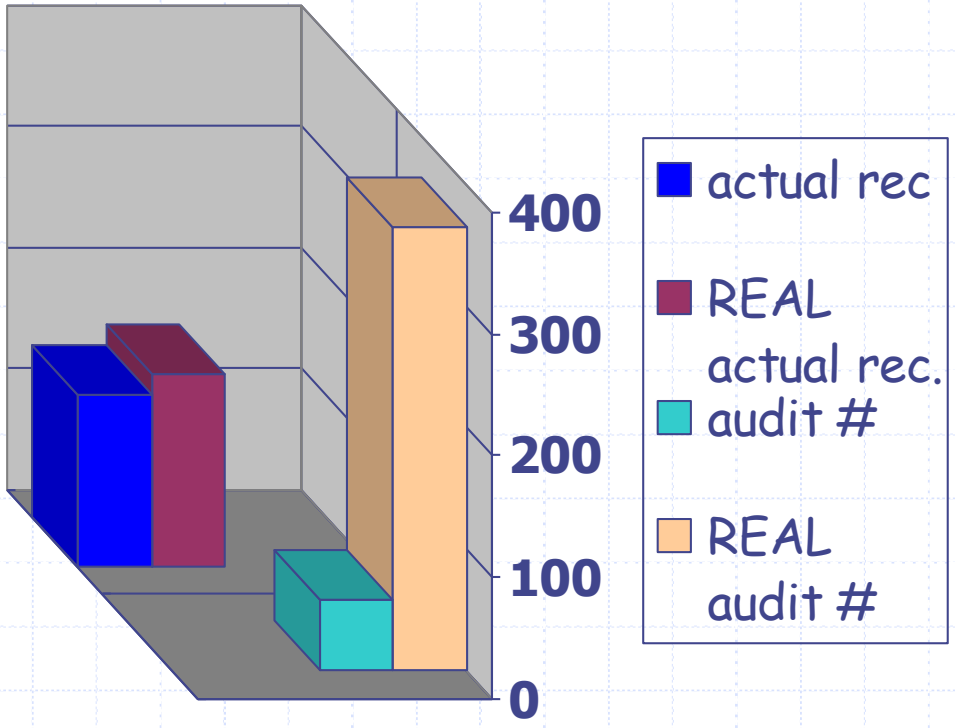
- audit # (di un dato classificatore): numero di tuple classificate come positive = # (FP ∪ TP)
- actual recovery: ammontare totale del recupero effettivo per tutte le tuple classificate come positive
- profitability: recupero effettivo medio per verifica
- relevance: rapporto tra profitability e l'errore di classificazione

Il caso REAL

- I Classificatori sono confrontati con l'intero test-set, cioè gli accertamenti veramente condotti.
- audit # (REAL) = 366
- actual recovery(REAL) = 159.6 M euro

Classificatore 1 (min FP)

- *misc. rate* = 22%
- *audit #* = 59 (11 FP)
- *actual rec.*= 141.7 Meuro
- *profitability* = 2.401



Classificatore 2 (min FN)

- *misc. rate* = 34%
- *audit #* = 188 (98 FP)
- *actual rec.*= 165.2 Meuro
- *profitability* = 0.878