Tecniche di data mining per l'analisi di processi di business

Candidato: Hind Chfouka
Tutori accademici: Andrea Corradini e Roberto Guanciale

Università di Pisa
Facoltà di Scienze Matematiche, Fisiche e Naturali
Corso di Laurea Triennale in Informatica

Tirocinio formativi da 12 CFU
24 Febbraio 2012

Il tirocinio

Contesto ed obiettivi

- Contestualizzazione
  - Progetto Rupos: Ricerca Usabilità Piattaforme Orientate ai Servizi
  - Collaborazione tra: Dipartimento di Informatica, Link.it, Hyperborea
  - Uso di tecniche di process mining per l'analisi dei processi
- Obiettivi
  - Sfruttare l'enorme quantità di dati negli event log
  - Integrare l'analisi dei processi con tecniche di data mining
  - Estendere una piattaforma di process mining con plugin che realizzano le nuove tecniche
  - Sperimentazione con dati generati artificialmente

Alcuni definizioni

- Che cos'è un processo di business?
- Cos'è il process mining?

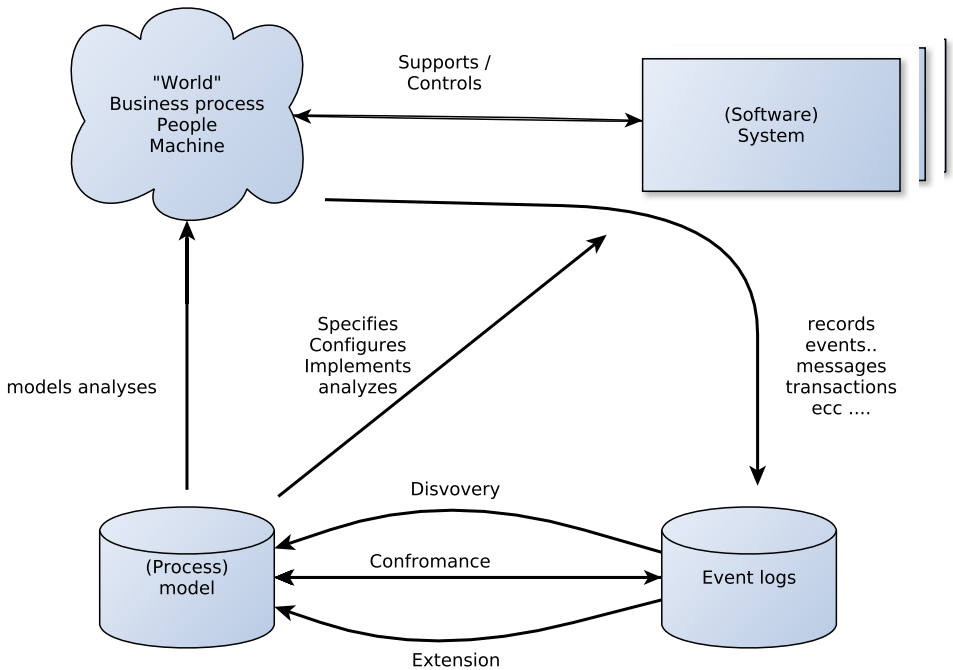
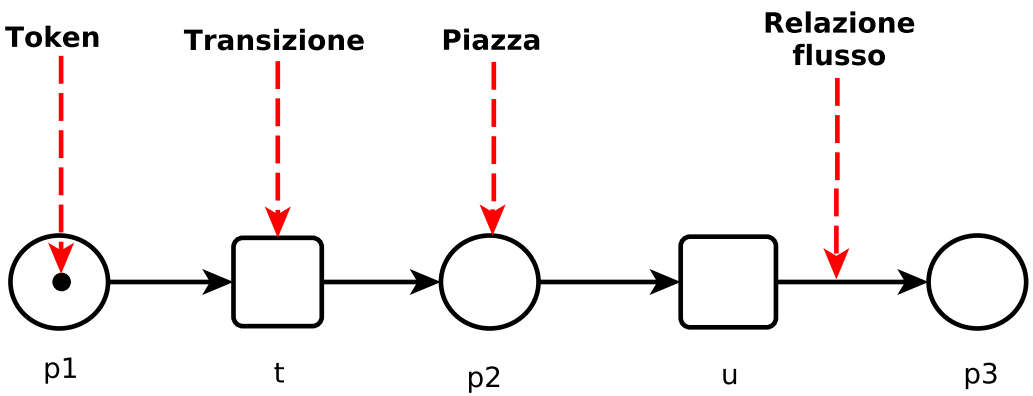


Figura: Process mining

Formalismi per i processi

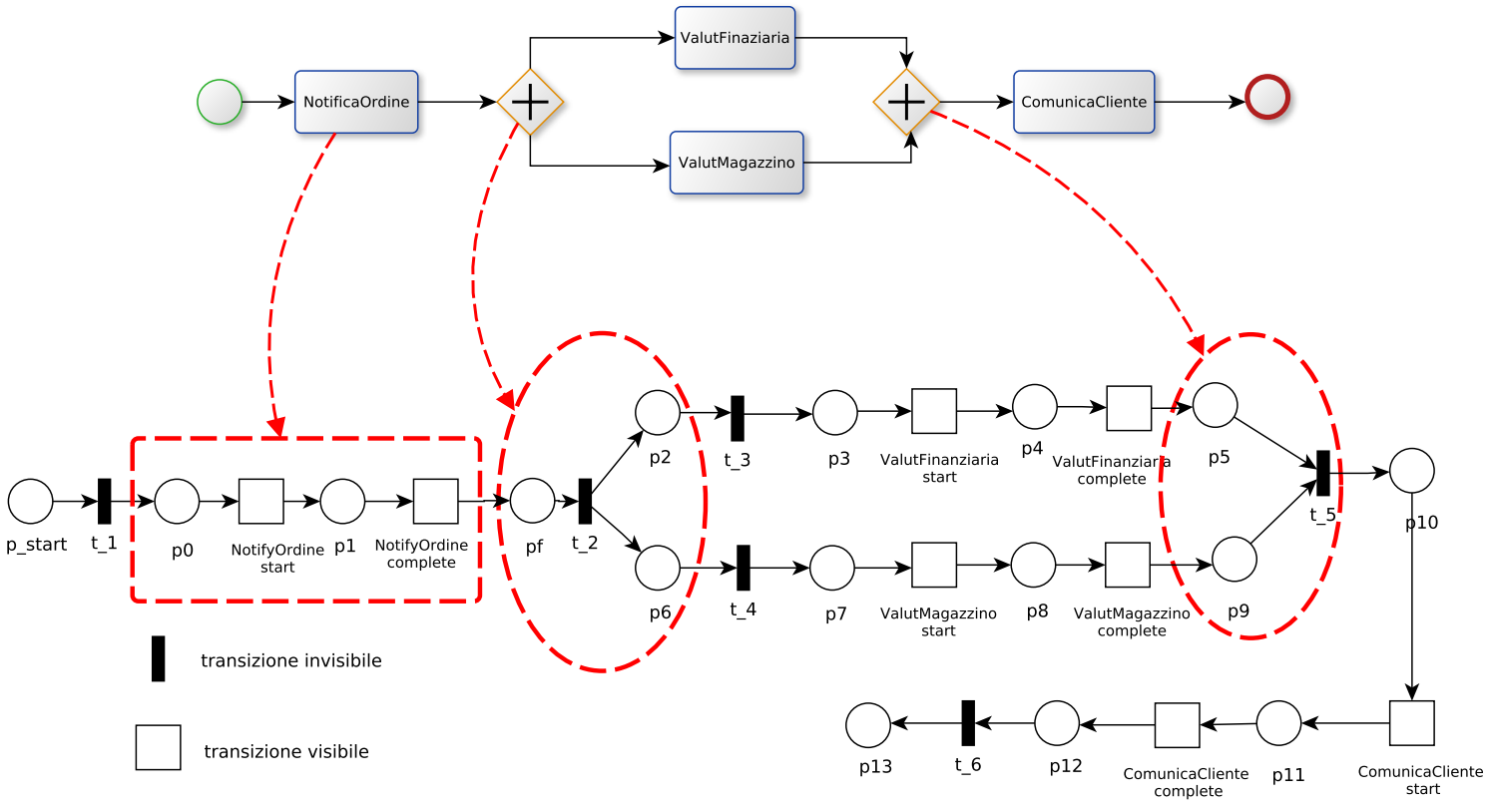
- BPMN: espressivo, ad alto livello, notazione grafica, intuitivo.
- Rete di Petri: modello matematico, formale, privo di ambiguità. Una rete di Petri è una quadrupla:

$$N = \langle P, T, F, M_0 \rangle$$



Mapping BPMN in rete di Petri

Esempio di modello di processo BPMN e trasformazione in rete di Petri equivalente.



Analisi basata su reti di Petri

Concetti preliminari

- Evento $e = (a, t)$, unità base di event log
- Traccia: sequenza finita di eventi $T[1], ..., T[n]$ ordinati sul timestamp. Rappresenta un'istanza
- Eventi mappati sulle transizioni della rete

Algoritmo di analisi

- Log replay: processa le tracce di un log in modo non bloccante
  1. Parte con un token nella piazza iniziale della rete
  2. Estrae l'evento in testa al log
  3. Viene effettuato lo scatto della transizione corrispondente
     - se la transizione non è abilitata vengono creati dei token artificialmente e chiamati: **token mancanti**
- I risultati dell'algoritmo sono usati per dedurre conformance e performance delle istanze

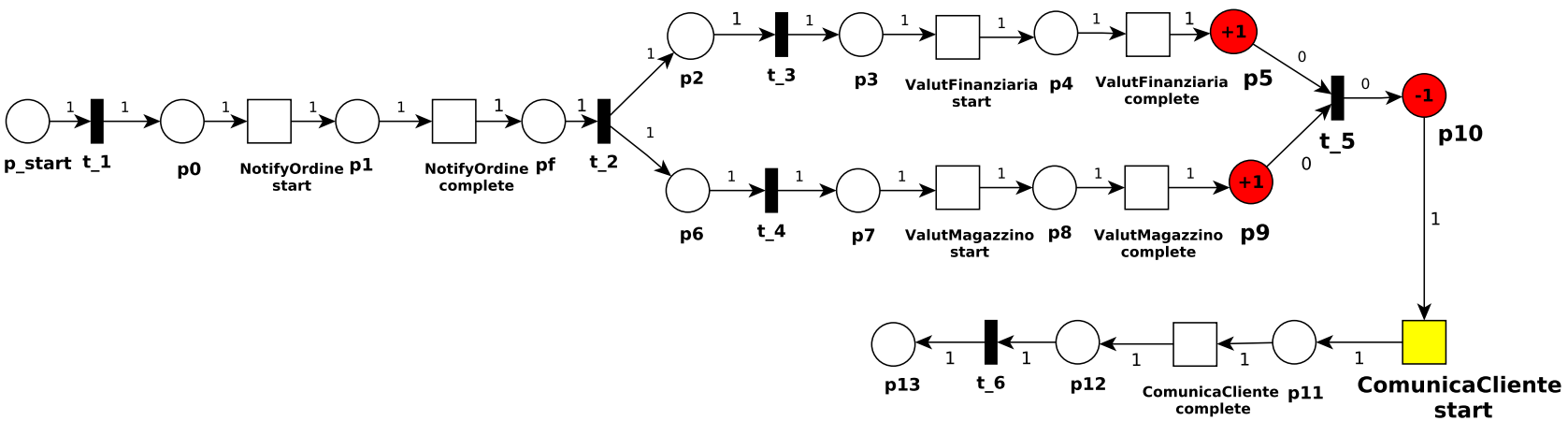
Analisi di conformance

Verificare se una traccia T soddisfa la rete di Petri PN.

Dai risultati del log replay...

- token mancanti: generati solo per eseguire transizioni visibili. Solo nelle piazze che hanno nel post_set una transizione visibile.

NotifyOrdine #1 start 02.12.2011 9:30:50	NotifyOrdine #1 complete 02.12.2011 10:30:00	ValutFinanziaria #1 start 02.12.2011 10:35:25	ValutFinanziaria #1 complete 02.12.2011 11:05:20	**ComunicaCliente #1 start 02.12.2011 11:40:20**	ComunicaCliente #1 complete 02.12.2011 12:10:20	ValutMagazzino #1 start 02.12.2011 11:40:20	ValutMagazzino #1 start 02.12.2011 11:40:20



Classificazione

- Dati: collezione di record.
- Record caratterizzato da $(\mathbf{x}, y)$: $\mathbf{x}$ insieme di attributi, y attributo target.
- I valori possibili di y sono noti.
- Classificazione: costruire una funzione f capace di associare ad ogni insieme di attributi $\mathbf{x}$ un valore y per l'attributo target. La funzione f è chiamata modello di classificazione o classificatore.
  - Algoritmo di apprendimento automatico per la classificazione.
    - Dati di input: training set, test set.

Alberi di decisione

- Strumento di classificazione con una struttura gerarchica:
  - radice e nodi interni: test condizionali su attributi.
  - arco: possibile risposta al test.
  - nodo terminale: etichetta di classe.
- Diversi algoritmi per generare alberi di decisione.
  E' stata scelta un'implementazione dell'algoritmo ricorsivo $C4.5$

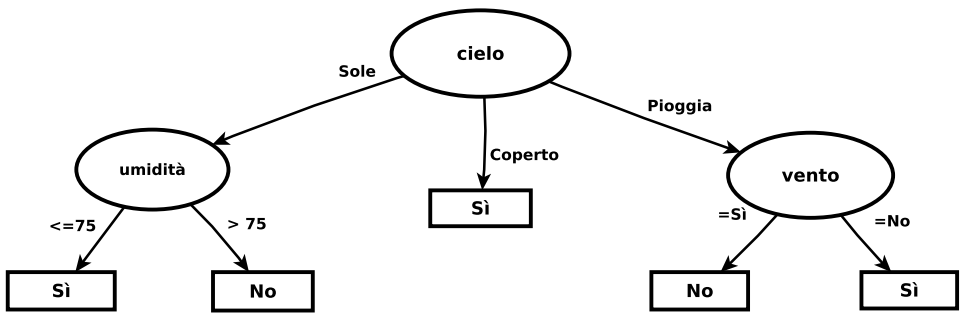


Figura: Albero decisionale: uscire a giocare?

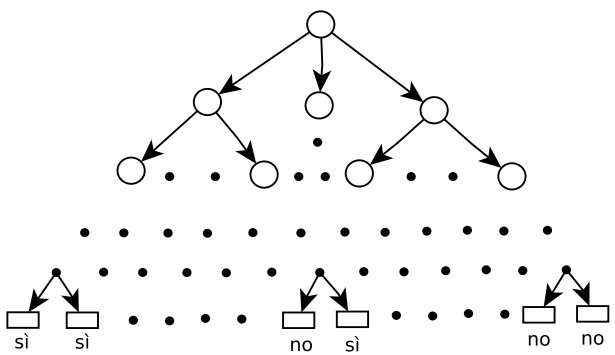
Conformance: approccio basato su classificazione

Idea di base

Sviluppare un approccio basato su classificazione in grado di individuare regole nei dati in corrispondenza delle quali si verificano errori di conformance.

L'attributo target: conformità o meno della traccia al modello (sì o no)

- Perché?
  - Scoprire le cause di errori ed adottare misure correttive.
  - Predire i casi di errore.



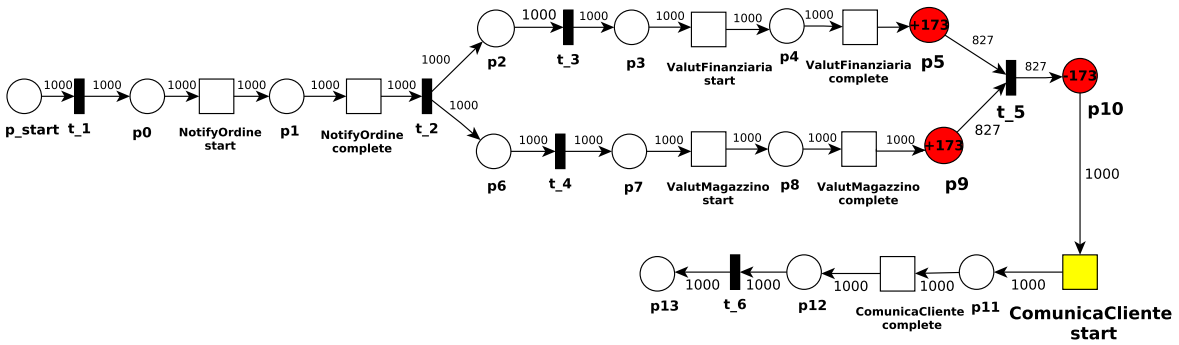
Caso di studio: Processo di vendita

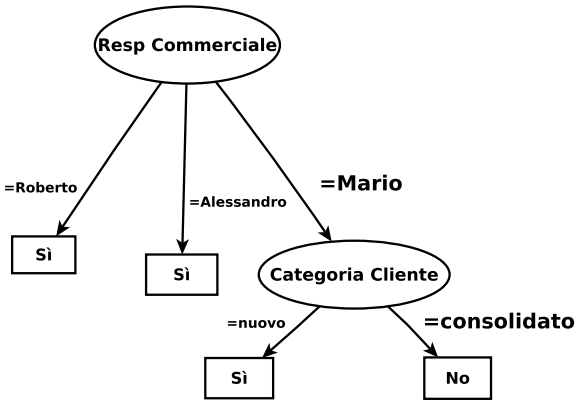
- Estrazione dei dati dall'event log

.....	**Categoria Cliente**	**Resp commerciale**	**Esito finanziario**	**Esito magazzino**	**Esito ordine**
.....	nuovo ..... consolidato	Mario ..... Roberto	positivo ..... negativo	positivo ..... negativo	confermato .... negato .....

- Esecuzione del log replay.
- Formulazione del problema di classificazione a partire dai dati e dal risultato di conformance.
  L'attributo target identifica la conformance di ogni traccia.
- Costruzione dell'albero di decisione

Interpretazione dei risultati





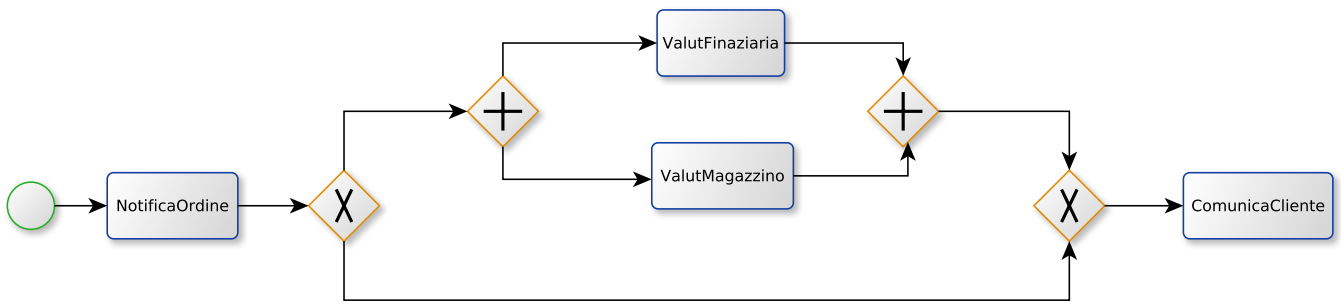
.....	Categoria Cliente	Resp commerciale	Esito finanziario	Esito magazzino	Esito ordine	Conformance

- Alcune istanze non sono conformi al modello: la comunicazione al cliente avviene prima della terminazione delle valutazioni.
- Gli ordini gestiti da Mario e fatti da clienti consolidati non rispettano la procedura.

Possibili misure correttive

A livello di processo

Riorganizzazione del processo aziendale: giudicare ragionevole saltare le fasi di valutazione per i clienti consolidati.



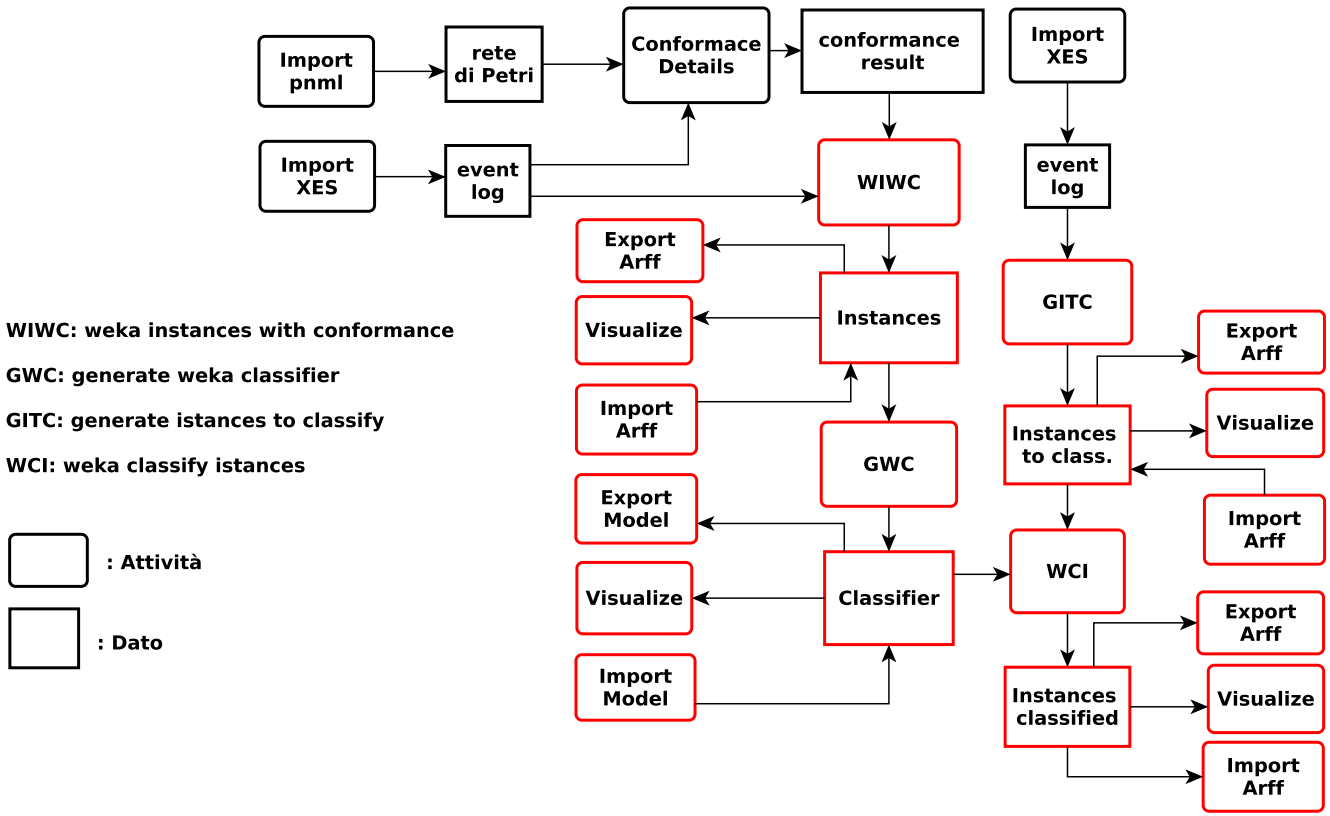
Predizione

Uso del classificatore in senso predittivo: prevedere i casi di non conformità con segnalazione al personale per evitare errori noti.

Framework di analisi

Tecnologie impiegate

- Process mining: ProM 6
- Data mining: Weka
- Linguaggio: Java



Conclusioni

Risultati

E' stato sperimentato un approccio basato sulla classificazione e sui risultati del log replay tramite:

- Integrazione della classificazione a servizio dell'analisi dei processi di business
- Estensione della piattaforma ProM 6 con i plugin realizzati
- Sperimentazione con prototopi di processo e dati sintetici

Sviluppi futuri

- Sperimentare l'approccio considerando processi di business che caratterizzano contesti organizzativi reali
- Estendere l'approccio prendendo in considerazione metriche di *performance*

Grazie per l'attenzione!

Analisi di performance

Metriche di performance

Durante il log replay, si sfruttano i timestamps per calcoloare per ogni piazza alcune metriche come il tempo di:

- sincronizzazione $tsc(p)$: intervallo tra arrivo del token in p e l'abilitazione di una transizione nel suo post_set. E' maggiore di zero solo se esiste una transizione nel pos_set dipendente anche da un'altra piazza (cioè quelle coinvolte nel modellare il Join Gatway).

Estensione alla performance

Siano p_1 e p_2 le piazze coinvolte nel modellare un Join Gatway. Si potrebbe considerare come attributo target del problema di classificazione:

- $max\{tsc(p_0), tsc(p_1)\}$ oppure $tsc(p_0) - tsc(p_1)$

Serve una fase di discretizzazione.

Valutare la performance

Matrice di confusione

La valutazione di un classificatore avviene sul numero di record del test set classificati correttamente.
Si usa una tabella detta: Matrice di confusione

Classe attuale	Classe predetta	
	Casse = 1	Classe = 0
Classe = 1	f_{11}	f_{10}
Classe = 0	f_{01}	f_{00}

f_{ij}: il numero dei record appartenenti alla classe i che sono stati classificati come appartenenti alla classe j

- $Accuratezza = \frac{Numero\ \ delle\ \ predizioni\ \ corrette}{Numero\ \ totale\ \ delle\ \ predizioni} = \frac{f_{11}+f_{00}}{f_{11}+f_{00}+f_{01}+f_{10}}$