Algoritmi di feature selection

Simone Rutigliano

Corso di Laurea in Informatica Magistrale

12 dicembre 2014

Outline

- **Java**
  - *PageRank*
  - *HITS*
  - *SALSA*
  - *ReConRank*
  - *SimRank*
  - *TripleRank*
  - *mRMR*
  - *PICSS*
- **RapidMiner**
  - *SHSEL*
    - *Information Gain*
    - *Correlation*
  - *GreedyTopDown*

PageRank

Implementazione del *Wrapper Model*:

- Utilizzare lo stesso algoritmo sia per la feature selection sia per la fase di raccomandazione
- Subset ottimizzato per la raccomandazione

HITS

Creazione del subset attraverso l'utilizzo dell'algoritmo di *Hyperlink-Induced Topic Search* basato sul *ranking* di risorse in base a due metriche:

- Hub
- Authority

Implementazioni trovate:

- `http://goo.gl/4pWAq4`
- `http://goo.gl/qSDXru`

SALSA

- Combinazione di HITS e PageRank
- Usa i punteggi di Hub e Autority
- Crea un grafo bipartito $G = (V_1 \cup V_2, E)$ dove
  - V_1 rappresenta il set degli Hub
  - V_2 rappresenta il set degli Autority
  - Una risorsa può essere contenuta sia in un set che nell'altro

Implementazione trovata:

- `http://goo.gl/DtHa4K`

ReConRank

Tratto dal paper [HHD06]

- Basato su due ranking:
  - **ResourceRank** : Associa uno score basato sul PageRank alle risorse del grafo RDF
  - **ContextRank** : Permette di inglobare la provenienza del contenuto semantico nel calcolo del ranking
- Computazione molto onerosa
- Implementazioni trovate
  - `http://goo.gl/PnZfNc`
  - `http://goo.gl/oCwQWe`

SimRank

Tratto dal paper [JW02] Algoritmo per il calcolo di similarità tra due nodi all'interno di un grafo G

- Esegue un random walk con ripartenza da un nodo fissato u su un grafo k-partito
- Gli score risultati misureranno la similarità tra il nodo u e tutti gli altri nodi del grafo

Implementazione trovata:

- `http://goo.gl/9cLDda`

TripleRank

Tratto dal paper [FSSS09]

- Consiste in una generalizzazione di HITS nel contesto dei Linked Data
- Permette di valutare al meglio le proprietà delle entità e di filtrare le relazioni semantiche dell'entità stessa presente nella linked data

Implementazione trovata:

- `http://goo.gl/Pb3vEr` (Richiede l'utilizzo di Matlab)

mRMR

Tratto dall'articolo [PLD05] e approfondito in [Rut14]

- Consiste nel calcolo della
  - **minima ridondanza** tra le features
  - **massima rilevanza** delle features con la classe target

  Implementazione trovata:
  - `http://goo.gl/YQUx1s`

PICSS

Trattato nella PhD Thesis di Meymandpour e negli articoli correlati [MD14] e [MD13]

- Tecnica di ranker ottenuta combinando:
  - **P***artitioned* **I***nformation* **C***ontent* : Seleziona una partizione della LOD in base al contesto da analizzare
  - **S***emantic* **S***imilarity measure* : Pesa gli archi tra feature in base all'information content che quel predicato apporta all'entità (Più viene utilizzato quel predicato meno apporto informativo conterrà)
- Non sono state trovate implementazioni di questo approccio

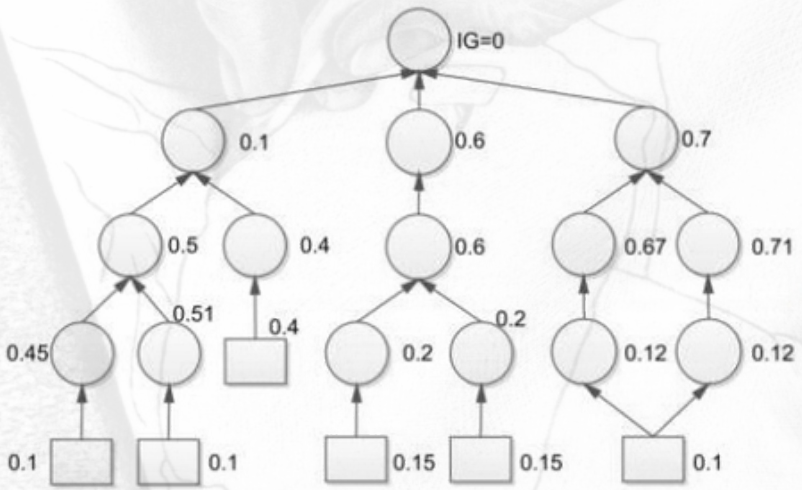
RapidMiner Linked Open Data Extension

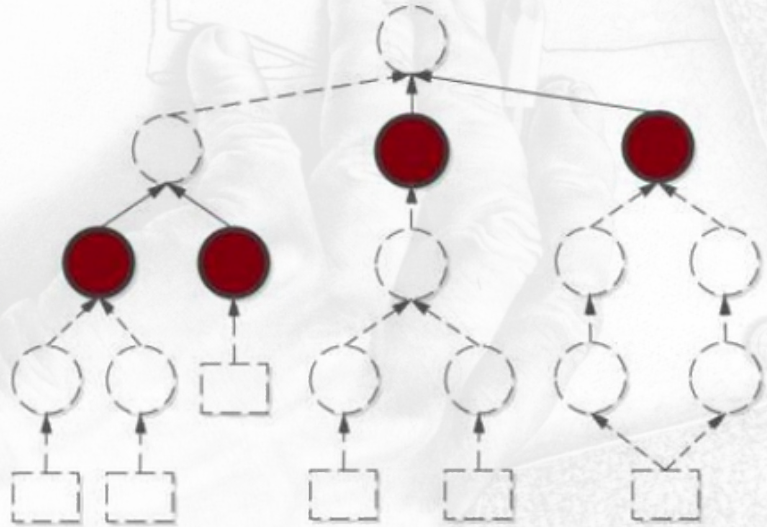
L'estensione per RapidMiner inerente i LOD sviluppata dalla University of Mannaheim [1] permette di utilizzare i seguenti algoritmi per la feature selection sui Linked Open Data:

- Greedy Top Down
- TSEL tramite Information Gain
- SHSEL tramite Information Gain

[1] Sito di riferimento `http://goo.gl/uoUx1k`

Optimal Feature Selection - Example

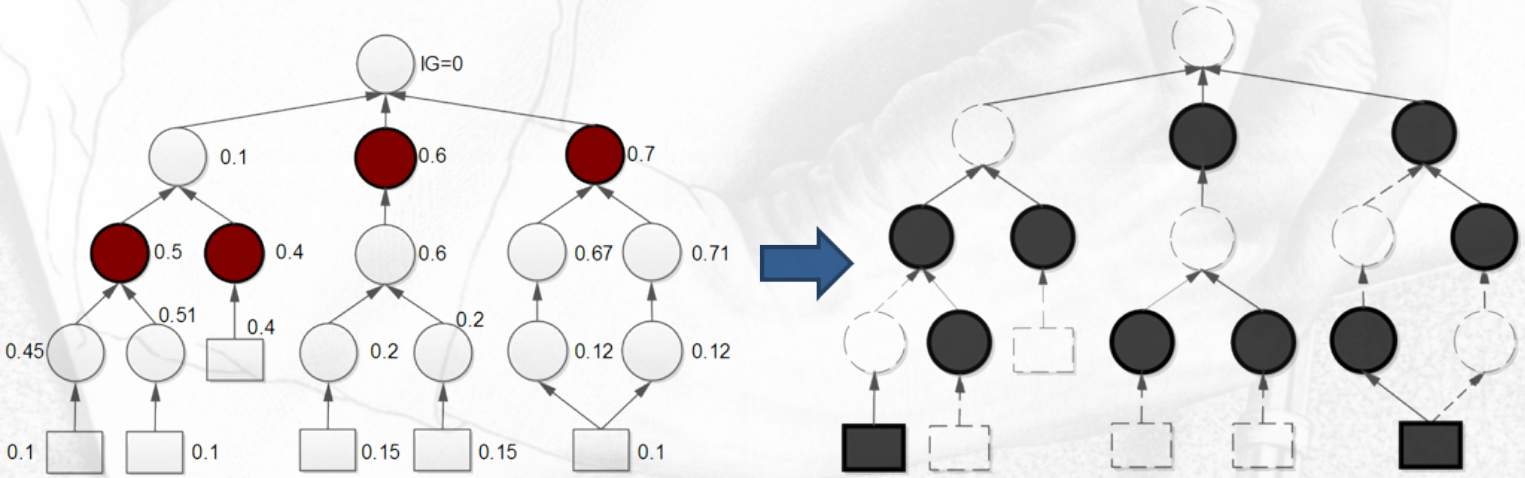




Greedy Top Down

Strategia di ricerca Greedy di tipo top down per la feature selection

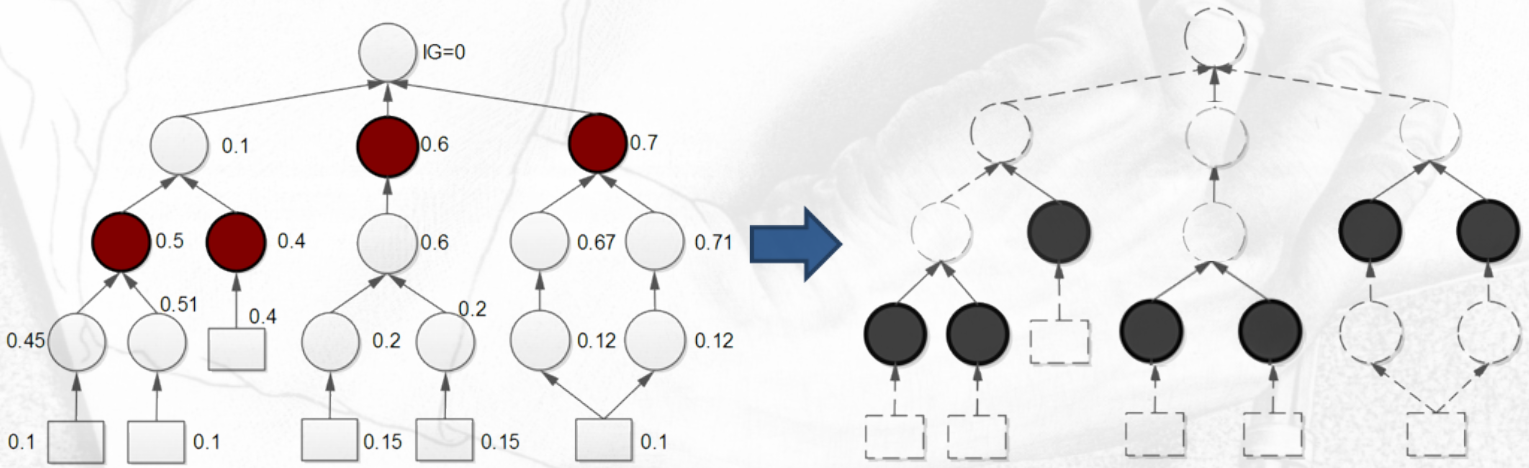
- Seleziona i nodi più rappresentativi da diversi livelli della gerarchia



TSEL - Information Gain

Tree-based feature selection tratto da [JM13]

- Seleziona le feature più rappresentative da ogni ramo della gerarchia



SHSEL - Information Gain ...

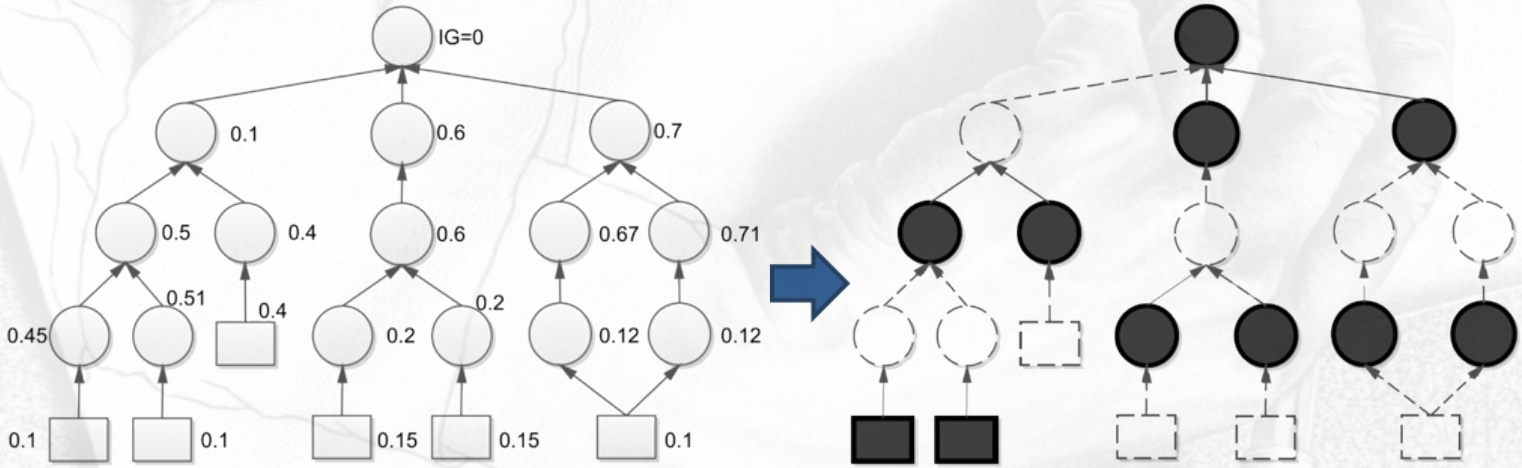
Descritto dal paper [RP14] e nel corrispettivo sito [2]

- Identifica le feature con rilevanza simile e seleziona le feature astratte migliori (quelle con livello gerarchico più alto, senza perdere potere predittivo)
- La misura di similarità applicata sarà l'information gain
- L'approccio prevede due fasi:
  - Selezione iniziale
  - Pruning

[2] http://goo.gl/NNlnuE

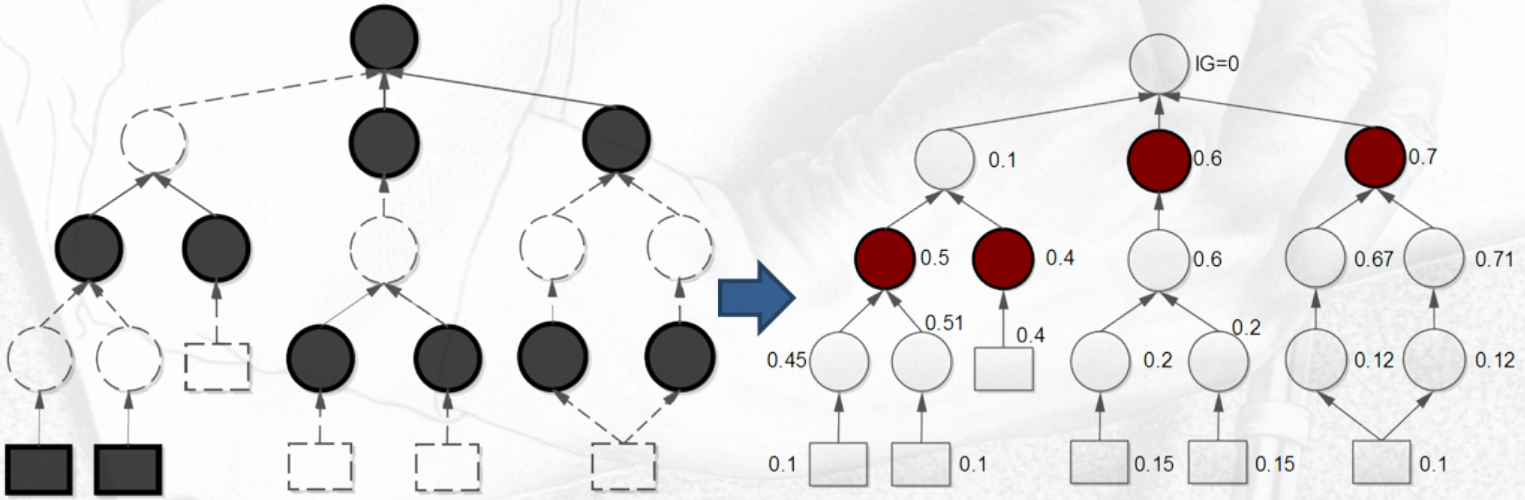
. . . SHSEL - Information Gain . . .

- Nella prima fase, in ogni ramo della gerarchia verranno identificati e filtrati i set di nodi con rilevanza simile



...SHSEL - Information Gain

- Nella seconda fase, si proveranno a selezionare solo le feature più rilevanti dal subset ottenuto nella fase precedente
- La metrica utilizzata sarà sempre il valore dell'information gain



References I

Thomas Franz, Antje Schultz, Sergej Sizov, and Steffen Staab.

Triplerank: Ranking semantic web data by tensor decomposition.
In *Proceedings of the 8th International Semantic Web Conference*, ISWC '09, pages 213–228, Berlin, Heidelberg, 2009. Springer-Verlag.

References II


Aidan Hogan, Andreas Harth, and Stefan Decker.
Reconrank: A scalable ranking method for semantic web data with context.
In *In 2nd Workshop on Scalable Semantic Web Knowledge Base Systems*, 2006.


Yoonjae Jeong and Sung-Hyon Myaeng.
Feature selection using a semantic hierarchy for event recognition and type classification.
In *Proceedings of the 6th International Joint Conference on Natural Language Processing*, pages 136–144. Asian Federation of Natural Language Processing, 2013.

References III


Glen Jeh and Jennifer Widom.
Simrank: A measure of structural-context similarity.
In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '02, pages 538–543, New York, NY, USA, 2002. ACM.


Rouzbeh Meymandpour and Joseph G. Davis.
Ranking universities using linked open data.
In Christian Bizer, Tom Heath, Tim Berners-Lee, Michael Hausenblas, and Sören Auer, editors, *LDOW*, volume 996 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2013.

References IV


Rouzbeh Meymandpour and Joseph G. Davis.
Lodify: A hybrid recommender system based on linked open data.
In *11th Extended Semantic Web Conference (ESWC 2014)*, Crete, Greece, 2014.


Hanchuan Peng, Fuhui Long, and Chris Ding.
Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy.
IEEE Transactions on Pattern Analysis and Machine Intelligence, 27:1226–1238, 2005.

References V


Petar Ristoski and Heiko Paulheim.
Feature selection in hierarchical feature spaces.
In *Discovery Science*, pages 288–300. Springer, 2014.


Simone Rutigliano.
mrmr slides.
https://github.com/Simoruty/mRMR-slides, 2014.