Algoritmi di feature selection

Simone Rutigliano

Corso di Laurea in Informatica Magistrale

12 dicembre 2014



Outline

PageRank
HITS
SALSA
ReConRank
SimRank
TripleRank
mRMR
PICSS

 $\textbf{RapidMiner} \begin{cases} \textit{SHSEL} & \textit{Information Gain} \\ \textit{Correlation} \end{cases}$ GreedyTopDown



PageRank HITS SALSA ReConRank SimRank TripleRank mRMR PICSS

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





PageRank HITS SALSA ReConRank SimRank TripleRank mRMR PICSS

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
 - V₁ rappresenta il set degli Hub
 - V₂ 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





PageRank HITS SALSA ReConRank SimRank TripleRank mRMR PICSS

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)



PageRank HITS SALSA ReConRank SimRank TripleRank mRMR PICSS

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]

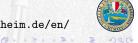
- Tecnica di ranker ottenuta combinando:
 - Partitioned Information Content : Seleziona una partizione della LOD in base al contesto da analizzare
 - Semantic Similarity measure: Pesa gli archi tra feature in base all'information content che quel predicato apporta all'entità (Più viene utilizzato quel 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 :

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



¹Sito di riferimento http://dws.informatik.uni-mannheim.de/en/research/rapidminer-lod-extension/

SHSEL - Information Gain

University of Mannheim



TSEL - Information Gain

University of Mannheim



Greedy Top Down

University of Mannheim



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.



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.



References III

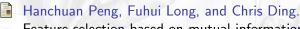


Rouzbeh Meymandpour and Joseph G. Davis.
Lodify: A hybrid recommender system based on linked open data.

In 11th Extended Semantic Web Conference (ESWC 2014), Crete, Greece, 2014.



References IV



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.

Simone Rutigliano.

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

