Documentazione Data Mining

Simone Rutigliano

22 ottobre 2013

Indice

Elenco delle figure

Elenco delle tabelle

1 Introduzione

Nell'era dell'information Overload, dove giornalmente si producono quantità considerevoli di dati, memorizzati su opportuni database aziendali e non, potrebbe risultare utile utilizzare degli strumenti in grado produrre automaticamente della conoscenza a partire da questa mole di dati. Una metodologia propensa a fare ciò, è la metodologia KDD (*Knowledge Discovery Databases*). La prima definizione da attribuire a questa metodologia è stata quella di:

Intero ***processo*** *di estrazione di conoscenza, dalla raccolta e pre-processing dei dati, fino alla interpretazione dei risultati(1)*

Successivamente Fayyad et al.,hanno raffinato tale definizione trasformandola in:

Knowledge discovery is the nontrivial extraction of implicit, previously unknown, and potentially useful information from data. (2) Definendo quindi il processo KDD come un processo in grado di estrarre dai dati delle informazioni non banali, sconosciute e potenzialmente utili. Il processo KDD si articola in diverse fasi cosi come mostrato in figura 1.

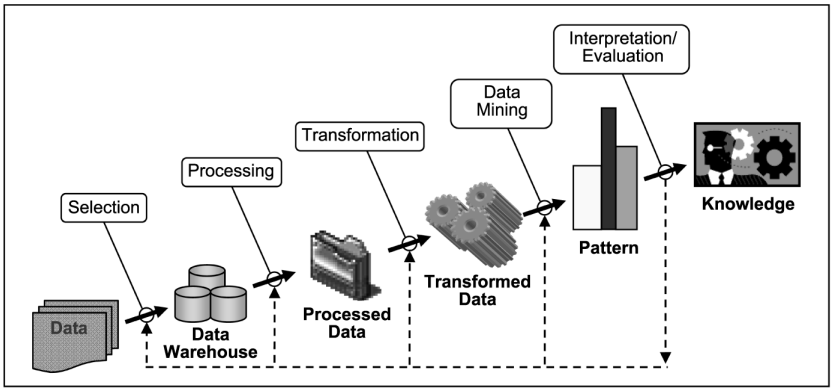

Figura 1: Processo KDD

Il Data mining rappresenta la fase principale del processo KDD, il cui compito può consistere o nell'adattare un modello esistente ai dati a disposizione, oppure nel determinare dei possibili pattern ricorrenti tra i dati osservati utilizzando o delle tecniche di machine learning oppure delle tecniche statistiche.

2 CRISP-DM

Per l'applicazione del processo di KDD si seguirà il modello del CRISP-DM (***CR****oss* ***I****ndustry* ***S****tandard* ***P****rocess for* ***D****ata* ***M****ining*) (4) , in quanto, tale

modello, risulta essere lo standard riconosciuto a livello industriale per la conduzione dei processi di KDD. Il CRISP-DM si compone di sei fasi il cui ordine non è prestabilito in modo vincolante ma può variare da applicazione ad applicazione. Tipicamente le fasi di cui si compone il modello, vengono eseguite come mostrato in figura 2.

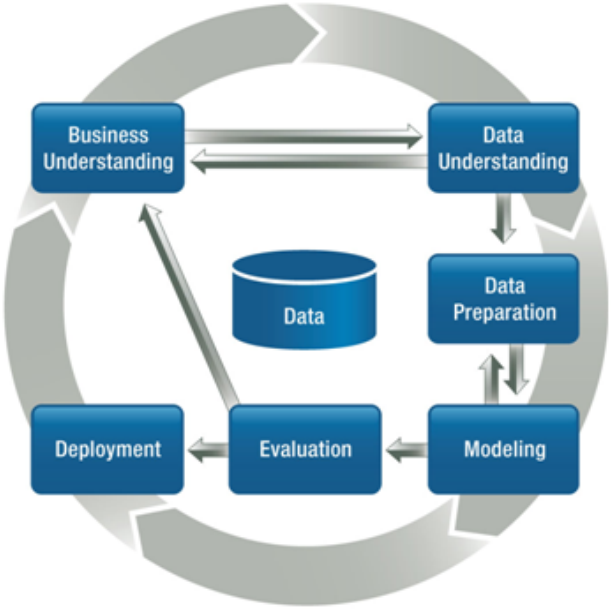

Figura 2: CRISP-DM

Di seguito verranno esaminate le singole fasi in maniera più approfondita e per ognuna di essa, verrà illustrato l'utilizzo fatto all'interno del contesto del progetto preso in esame.

3 Business Understanding

Questa fase si focalizza sulla individuazione degli obiettivi e i requisiti del progetto dal punto di vista del business.

3.1 Background

Il contesto di

Il problema delle richieste consumati (pubblicità) e-mail , di solito con contenuti discutibili , è ben nota e la posizione del ricevitore probabilmente abbastanza disperata . Nonostante appropriate norme legislative persecuzione drastica in senso giuridico è difficilmente praticabile . Per le persone con numerosi contatti esterni tramite e-mail , e quindi una corrispondente distribuzione del loro indirizzo , la situazione può anche rappresentare un modo che, non solo i messaggi di posta elettronica desiderati ma anche molte volte le e-mail più promozionali (spam) nella casella di posta al giorno .

Il ricevitore è ora di fronte il problema per lui per separare le email importanti dai messaggi di spam che non sono sempre evidenti a prima vista , in quanto tale , - un processo noioso e che richiede tempo .

Il primo desiderio di automatizzare il rilevamento di spam e la loro selezione , ha portato nel corso degli anni a vari progetti di software e di prodotti commerciali con più o meno artificiale intelligenza e una corrispondente apertura alare di efficienza . Uno dei più noti posta - filtro più è l' open source software SpamAssassin (tm), che deve fornire la loro idoneità per la pratica in molte università e aziende per il test . Fondamento essenziale di un filtro di posta buona è la sua capacità di classificare i messaggi di posta elettronica sulla base di diverse caratteristiche controllate e riconosciuto e, quindi, per rendere l'assegnazione di spam o non-spam con una corrispondente alta probabilità. La qualità di un filtro di posta è e cade con la qualità del suo algoritmo di classificazione.

Die Problematik unverlangt zugesendeter (Werbe-) E-Mails, mit zumeist fraglichem Inhalt, ist weithin bekannt und die Lage der Empfänger wohl recht aussichtslos. Trotz entsprechender Gesetzes- regelungen ist eine drastische Verfolgung im juristischen Sinne kaum praktikabel. Für Personen mit zahlreichen Außenkontakten per E-Mail und damit einer entsprechenden Verbreitung ihrer Adresse kann sich die Situation sogar so darstellen, dass pro Tag nicht nur die gewünschten E-Mails sondern auch das Mehrfache an Werbe-E-Mails (Spam) im Postfach liegen.

Der Empfänger steht nun vor dem Problem, die für ihn wichtigen E-Mails von den Spam-Mails, die ja nicht immer auf den ersten Blick als solche zu erkennen sind, zu trennen - ein lästiges und zeitraubendes Verfahren.

Der frühe Wunsch, die Erkennung von Spam-Mails und deren Auslese zu automatisieren, führte über die letzten Jahre zu verschiedenen Software-Projekten und auch kommerziellen Produkten mit mehr oder weniger künstlicher Intelligenz und einer entsprechenden Spann- weite an Effizienz. Einer der wohl bekanntesten Mail-Filter ist die Open-Source Software SpamAssassin(tm), der ihre Praxis- tauglichkeit schon in vielen Hochschulen und Unternehmen unter Beweis stellen muss. Wichtiges Kernstück eines guten Mail-Filters ist seine Fähigkeit, E-Mails auf Grund verschiedenster kontrollierter und erfasster Merkmale zu klassifizieren und somit mit einer entsprechend hohen Wahrscheinlichkeit die Zuordnung Spam oder Nicht-Spam zu treffen. Die Qualität eines Mail-Filters steht und fällt somit mit der Güte seines Klassifikationsalgorithmus.

3.1.1 Risorse

La principale risorsa utilizzata è l'hardware del sistema utilizzato per eseguire l'algoritmo di data mining, in particolar modo, un sistema Windows con un Quad-Core Intel 2.20 GHz e 4 GB di Ram. Il tool di data mining scelto è WEKA (Waikato Environment for Knowledge Analysis) (3): una popolare suite di software per il machine learning scritto in Java e sviluppato nell'Università di Waikato (Nuova Zelanda). I motivi che hanno portato alla scelta di questa suite, ricade nel fatto che tale software è :

- Liberamente scaricabile dal sito [1];
- Portabile, in quanto totalmente implementato utilizzando java;
- Ampia gamma di tecniche di preprocessing e modellazione dei dati;
- Facile da usare grazie alla GUI;

Per quanto riguarda il personale umano, l'unica risorsa umana interpellata nella sperimentazione è lo sperimentatore stesso.

3.1.2 Vincoli

Excluding the only assumption that data are freely available their accessibility is ensured by the fact that they were published on-line there are no other requirements, assumptions and constraints such as security, legal or privacy issues, or budget and resources constraints.

Non sono presenti vincoli temporali o problemi legali che devono essere considerati durante il processo di KDD.

[1] Weka site: `http://www.cs.waikato.ac.nz/ml/weka/downloading.html`

3.1.3 Assunzioni

Si assume che i dati di cui si intende disporre siano liberamente accessibili e che non siano falsi o errati.

3.2 Obiettivi di Business

Attraverso l'utilizzo delle email ricevute dagli utenti nel corso degli anni, Using the existing characteristics of a customer's initial order, such as order quantity per type of goods, title and delivery weight, the task is to indentify those customers that could be encouraged to buy again if a voucher worth ¿ 5.00 is send to them. Obviously, the customers who receive a voucher should be those who could not have decided to re-order by themselves, as not to incur a loss. The aim is therefore sending vouchers only to selected customers.

3.2.1 Task di Data Mining

3.3 Criteri di successo

3.4 Glossario dei termini

3.5 Analisi Costi-Benefici

3.6 Piano di Progetto

4 Data Understanding

4.1 Raccolta dei dati

4.2 Descrizione dei dati

4.3 Verifica della qualità dei dati

4.4 Esporazione dei dati

5 Data Preparation

5.1 Criteri di Inclusione/Esclusione dei dati

5.2 Selezione dei dati

5.3 Campionamento

5.4 Feature Selection

5.5 Data Cleaning

5.6 Construct Data

5.7 Integrate Data

5.8 Format Data

6 Modeling

6.1 Tecnica di Modeling

6.2 Rappresentazione del Modello

6.3 Valutazione del Modello

6.4 Ricerca

6.5 Test Design

6.6 Costruzione del Modello

6.7 Valutazione del Modello

7 Evaluation

7.1 Valutazione rispetto agli obiettivi di business

7.2 Raccomandazioni per revisioni future

8 Deployment

Riferimenti bibliografici

[1] U. M. Fayyad and R. Uthurusamy, editors. *Proceedings of the First International Conference on Knowledge Discovery and Data Mining (KDD-95), Montreal, Canada, August 20-21, 1995*, 1995. AAAI Press. ISBN 0-929280-82-2.

[2] U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy. *Advances in Knowledge Discovery and Data Mining*. The MIT Press, Feb. 1996. ISBN 0262560976. URL `http://www.worldcat.org/isbn/0262560976`.

[3] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The weka data mining software: an update. *SIGKDD Explor. Newsl.*, 11(1):10–18, Nov. 2009. ISSN 1931-0145. doi: 10.1145/1656274.1656278. URL `http://doi.acm.org/10.1145/1656274.1656278`.

[4] R. Wirth and J. Hipp. Crisp-dm: Towards a standard process model for data mining. In *Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining*, pages 29–39. Citeseer, 2000.