Università degli Studi di Bari

FACOLTÀ DI SCIENZE MATEMATICHE, FISICHE E NATURALI
Corso di Laurea in Informatica Magistrale

DATA MINING

Documentazione progetto

Studente:
Simone Rutigliano

Anno Accademico 2012-2013

Indice

Elenco delle figure

Elenco delle tabelle

Capitolo 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 1.1.

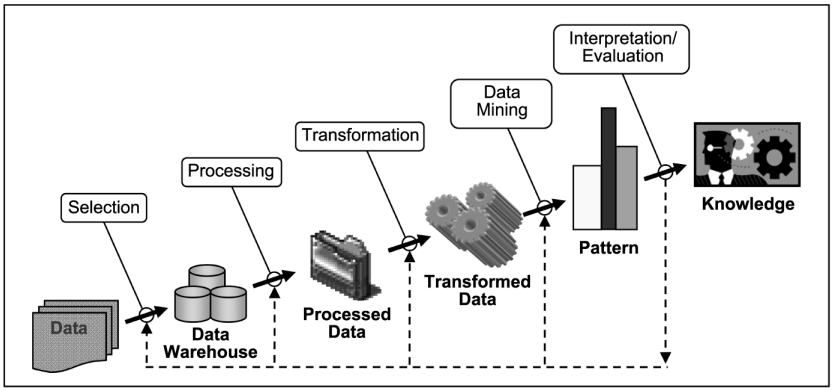

Figura 1.1: Processo KDD

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.

1.1 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*) (6) , 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 1.2.

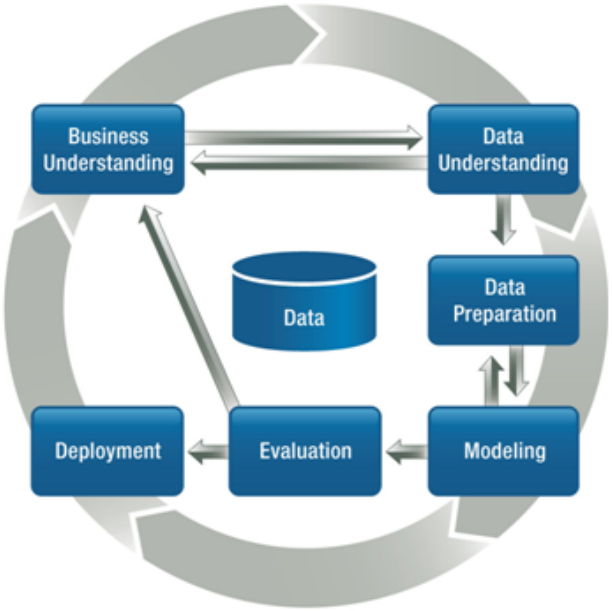

Figura 1.2: CRISP-DM

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.

Capitolo 2

Business Understanding

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

2.1 Background

Ogni giorno, vengono inviate circa 25 milioni di email indesiderate, chiamate anche email di spam. Tale cifra corrisponde a quasi il 10 % di tutte le email inviate nel mondo; inoltre, indagini svolte sull'argomento, rivelano che generalmente il 40 % delle email ricevute giornalmente dai dipendenti di molte imprese risultano essere email di spam, arrivando in alcuni casi, anche al 90 %. Queste percentuali risultano essere molto elevate principalmente a causa della facile diffusione delle proprie caselle di posta verso qualunque tipo di contatto, diventando cosi bersagli di messaggi promozionali di qualunque tipo. Questa inondazione di messaggi di spam quindi genera due problemi principali:

- Saturazione della propria casella di posta, anche se al giorno d'oggi si dispone di una capienza elevata;
- Perdita di tempo abbastanza considerevole da parte del ricevente nel filtrare queste email.

Nel corso degli anni, il desiderio di automatizzare il rilevamento e relativa selezione di queste email di spam, ha portato alla creazione e diffusione di numerosi progetti software e di prodotti commerciali in grado di filtrare lo spam in maniera tutto sommato efficiente; uno dei più diffusi è *SpamAssassin* [1], programma opensource rilasciato sotto licenza Apache 2.0. Si basa su regole

[1] `http://spamassassin.apache.org/`

di confronto del contesto, supporta anche regole basate su DNS, checksum e filtraggio statistico, inoltre supporta programmi esterni e database online. SpamAssassin è considerato uno dei filtri antispam più efficaci, specialmente se usato congiuntamente con un database antispam.(5)

2.1.1 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 i7 2.00 GHz e 4 GB di Ram. Il tool di data mining scelto è WEKA (**W**aikato **E**nvironment for **K**nowledge **A**nalysis) (3): una popolare suite di software per il machine learning scritta in Java e sviluppata nell'Università di Waikato (Nuova Zelanda); è stato deciso di utilizzare tale suite, in quanto il software porta con sè i seguenti vantaggi :

- Liberamente scaricabile dal sito [2];
- Portabile, in quanto totalmente implementato in java;
- Ampia gamma di tecniche di preprocessing e modellazione dei dati;
- Facile da usare grazie alla GUI;

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

2.1.2 Vincoli

Non sono presenti nè vincoli temporali, nè problemi legali legati alla diffusione del dataset in questione.

2.1.3 Assunzioni

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

2.2 Obiettivi di Business

Secondo quanto detto in precedenza, l'obiettivo di business di questo progetto consiste nell'individuazione delle email di spam attraverso l'utilizzo di tecniche di Data Mining.

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

2.2.1 Task di Data Mining

Il task da realizzare è di tipo *predittivo*, in particolar modo, sarà un task di *classificazione*. L'obiettivo sarà quindi quello di creare un classificatore che sia in grado di etichettare correttamente le nuove email come *spam* o *nospam* sulla base del training set dato in pasto al classificatore.

2.3 Criteri di successo

Il processo di KDD andrà a buon fine qualora siano rispettate due condizioni:

- Nel risultato del filtraggio, il numero di email di tipo non-spam contenute in questo risultato, deve essere limitato all' 1% della totalità;
- Minimizzare il numero di messaggi di spam che passeranno attraverso il filtro.

2.4 Glossario dei termini

Messaggi di spam: *Lo spamming, detto anche fare spam o spammare, è l'invio di messaggi indesiderati (generalmente commerciali). Può essere attuato attraverso qualunque sistema di comunicazione, ma il più usato è Internet, attraverso messaggi di posta elettronica, chat, tag board o forum.* (cit. Wikipedia)(4)

2.5 Analisi Costi-Benefici

Per quanto riguarda i costi inerenti al processo di KDD, l'unico costo che si avrà, sarà in termini di risorse temporali utilizzate per la realizzazione e relativa verifica dei risultati che il classificatore produrrà.

Invece, i benefici che si otterranno da questo processo, saranno quelli che andranno a sopperire a ciò che è stato detto in precedenza nel paragrafo 2.1.

2.6 Piano di Progetto

La durata stimata del progetto è pari a 2-3 settimane, la complessità dei database utilizzati influirà notevolmente sulla durata complessiva del progetto.

Le fasi da eseguire nel progetto sono quelle che sono fornite dal modello di processo CRISP-DM 1.1. Dal punto di vista dello sforzo che deve essere speso,

è spesso postulato che il 50-70 per cento del tempo e sforzi in un progetto di data mining viene utilizzato nella fase di preparazione dei dati e 20-30 percento in fase comprensione dati , mentre solo il 10-20 per cento viene speso in ciascuna delle fasi di comprensione di modellazione, di valutazione e di business. Il sospetto è quindi che il punto critico del progetto potrebbe essere la fase di preparazione dei dati che se povero comporterebbe la necessità di andare indietro nel processo e rivisitare i dati. Per quanto riguarda il tempo necessario a concludere il progetto, si prevede un massimo di due settimane.

2.7 Data Understanding

2.7.1 Raccolta dei dati

2.7.2 Descrizione dei dati

2.7.3 Verifica della qualità dei dati

2.7.4 Esportazione dei dati

2.8 Data Preparation

2.8.1 Criteri di Inclusione/Esclusione dei dati

2.8.2 Selezione dei dati

2.8.3 Campionamento

2.8.4 Feature Selection

2.8.5 Data Cleaning

2.8.6 Construct Data

2.8.7 Integrate Data

2.8.8 Format Data

2.9 Modeling

2.9.1 Tecnica di Modeling

2.9.2 Rappresentazione del Modello

2.9.3 Valutazione del Modello

2.9.4 Ricerca

2.9.5 Test Design

2.9.6 Costruzione del Modello

2.9.7 Valutazione del Modello

2.10 Evaluation

2.10.1 Valutazione rispetto agli obiettivi di business

2.10.2 Raccomandazioni per revisioni future

2.11 Deployment

Bibliografia

[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] Wikipedia. Spam — wikipedia, l'enciclopedia libera, 2013. URL `http://it.wikipedia.org/w/index.php?title=Spam&oldid=62087596`. [Online; in data 23-ottobre-2013].

[5] Wikipedia. Spamassassin — wikipedia, the free encyclopedia, 2013. URL `http://en.wikipedia.org/w/index.php?title=SpamAssassin&oldid=570651966`. [Online; accessed 22-October-2013].

[6] 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.