Electronic Logbook Data Mining

Giulio Angiani[1], Alberto Ferrari[1], Fosca Giannotti[3],
Agostino Poggi[1], Elena Salvatori[2]

[1] Dipartimento di Ingegneria e Architettura, Università di Parma
giulio.angiani@unipr.it alberto.ferrari@unipr.it agostino.poggi@unipr.it
[2] Dipartimento di Informatica, Università di Pisa
elena.salvatori@di.unipi.it
[3] Knowledge Discovery and Data Mining Laboratory, CNR
fosca.giannotti@isti.cnr.it

Sommario Situazioni di fallimento o di scarso successo nell'apprendimento scolastico sono un grave problema sociale ed è molto importante per i responsabili della formazione comprendere meglio il motivo per cui così tanti giovani non riescono a portare a termine i loro studi. La diffusione nella maggior parte delle scuole italiane della gestione elettronica dei registri genera una grande quantità di dati relativi alle attività didattiche: dalla frequenza scolastica degli studenti ai risultati ottenuti nelle singole prove di verifica, dalla tipologia di queste prove alla loro collocazione temporale nell'anno scolastico.
L'obiettivo del nostro progetto è quello di recuperare e analizzare con tecniche di Data Mining i dati provenienti dalle singole scuole per ottenere informazioni che tendano a segnalare e anticipare situazioni problematiche e in generale essere di ausilio al miglioramento degli obiettivi didattici.

Keywords: Educational Data Mining, Machine Learning

1 Il Progetto ELDM

Il progetto Electronic Logbook Data Mining si basa sull'applicazione di tecniche di Data Mining [3] [4] ai dati, opportunamente anonimizzati, presenti nei registri elettronici delle scuole medie superiori allo scopo di ricercare comportamenti e fenomeni sia ricorrenti che anomali che si possono verificare nel corso dell'anno scolastico.

A differenza delle normali metodologie statistiche, che analizzano i dati per verificare ipotesi di correlazione fra eventi diversi, il nostro studio sarà "data-driven" come viene normalmente indicato in letteratura scientifica l'approccio utilizzato nell'ambito dell'Educational Data Mining.

Educational Data Mining (EDM) [5] è un'area di ricerca interdisciplinare emergente che si è posta all'attenzione della comunità scientifica negli ultimi anni e si occupa dello sviluppo di metodi per esplorare i dati originari di un contesto educativo. L'analisi dei dati utilizzando approcci computazionali ha il

fine di studiare le questioni educative, ottenere informazioni relative all'attività degli studenti, ai loro risultati e alle loro modalità di apprendimento.

Per ottenere valide informazioni dai dati educativi vengono utilizzate metodologie statistiche, di Machine Learning e Data Mining per operare su grandi quantità di dati al fine di cercare di comprendere meglio gli studenti e le loro modalità di apprendimento. [1]

Varie tecniche di classificazione permettono di predire un insieme di risultati (nel nostro caso successo o insuccesso scolastico) basandosi su un insieme di valori derivanti da varie caratteristiche (features) che nel nostro caso sono individuate dai dati presenti nei registri scolastici.

Si tratta inoltre di "osservare" i dati grezzi per confrontarli con quelli aggregati, frutto di analisi statistiche sui risultati da prove ministeriali, per individuare conferme o scostamenti di valutazione utili per un possibile intervento mirato nelle attività di supporto al recupero di situazioni problematiche. [2]

L'utilizzo di metodologie di Machine Learning sui dati permette inoltre un'analisi senza alcun pregiudizio né idea preconcetta cercando al loro interno la presenza o meno di "pattern", ovvero di comportamenti, non necessariamente noti a priori.

Un'analisi di questo tipo, per avere valore scientifico reale, necessita di una grande mole di dati (approccio Big Data) e per questo motivo, dopo un primo studio sui dati di un insieme di scuole pilota, è stato deciso di allargare il progetto richiedendo la partecipazione al progetto di altre scuole medie superiori.

La gestione dei dati complessivi opportunamente anonimizzati potrà permettere una visione globale mentre alle singole scuole potranno essere restituite informazioni specifiche più dettagliate che permettano un'analisi locale più approfondita. L'insieme dei dati che utilizziamo ci permette di seguire passo-passo l'evoluzione della situazione a livello del singolo studente durante le attività scolastiche. Oltre alla classe frequentata, l'indirizzo di studio e la scuola di appartenenza, vengono registrati i dati relativi alla frequenza, alle valutazioni ricevute durante l'anno scolastico e ai voti di fine periodo.

Per rispettare la privacy e l'anonimato nel dataset non sono presenti dati personali di studenti, docenti e personale scolastico e le informazioni sono gestite ai sensi delle vigenti normative sul *Trattamento dei Dati Personali ai fine di ricerca e statistica* [6].

2 Ambiti di ricerca

Il recupero e la successiva analisi dei dati sono finalizzati a vari ambiti di ricerca.

2.1 Identificazione di comportamenti ricorrenti nello storico delle misurazioni degli studenti

L'utilizzo dell'approccio Big Data permette di far emergere, se esistono, comportamenti ricorrenti presenti nei dati analizzati. Questo tipo di ricerca può individuare pattern non noti a priori che possono fornire importanti informazioni spesso non intercettabili con analisi basate su un insieme ristretto di studenti.

L'obiettivo è far emergere best practices presenti nell'offerta formativa delle scuole individuando, dal raffronto dei risultati che emergeranno dallo studio dei dati, ambiti di miglioramento nelle azioni didattiche di recupero e di valorizzazione delle competenze.

2.2 Visualizzazione di dati educativi

Data la natura multidimensionale dei dati educativi, è necessario uno studio sulle modalità di selezione di tali dimensioni per ottimizzare la visualizzazione dei risultati e, a beneficio degli enti interessati, massimizzare il contenuto informativo presente nei registri delle scuole.

Tecniche di classificazione permettono di predire un insieme di valori (nel nostro caso successo o insuccesso scolastico) basandosi su un insieme di valori derivanti da varie caratteristiche (features) che nel nostro caso sono individuate dai dati presenti nei registri scolastici.

2.3 Relazione tra misurazioni della scuola e risultati delle prove ministeriali

Una seconda fase dello studio potrà prendere in esame le misurazioni fornite dalle scuole aderenti al progetto e le metterà in relazione ai risultati delle prove INVALSI. Questa analisi verrà effettuata all'interno dello stesso anno scolastico per cercare eventuali correlazioni fra tali dati. L'obiettivo è di individuare l'esistenza o meno di scostamenti evidenti fra la modalità di valutazione utilizzata nelle scuole e quella dei test ministeriali.

2.4 Predire insuccessi

A partire dai dati delle misurazioni di un anno scolastico e, se presenti, dai risultati dei test INVALSI si cercheranno correlazioni con l'andamento degli studenti nell'anno scolastico successivo.

L'obiettivo, nel caso siano presenti legami evidenti, è di predire delle difficoltà già nella prima parte dell'anno scolastico. Questo permetterebbe alle scuole di intervenire con strumenti di recupero con notevole anticipo rispetto alle pratiche consuete che sono tipicamente guidate dai risultati e quindi a valle dell'accertamento di una situazione problematica.

In letteratura internazionale esistono già degli studi che predicono con alta affidabilità l'insuccesso scolastico utilizzando un Dataset di 15 variabili; lo studio mostra l'alto valore predittivo delle valutazioni in lingua, lingua straniera e matematica [2].

3 Presentazione dei risultati

Al termine del progetto prevediamo di presentare i risultati della ricerca con due focus principali: uno che abbia una visione *globale* ed uno *locale*.

L'analisi *globale* permetterà di evidenziare analogie e differenze delle misurazioni didattiche nel territorio interessato dalle scuole partner del progetto. L'obiettivo dichiarato è di individuare i più evidenti gradienti di miglioramento relativi a diversi periodi dell'anno scolastico. Questo potrà poi permettere di ricercare, negli istituti con migliori risultati, le tecniche e le pratiche che hanno portato a tali risultati in modo da poter "contaminare" le altre scuole nell'ottica di un miglioramento continuo dell'offerta formativa. Particolare attenzione sarà data nell'evidenziare eventuali anomalie e comportamenti imprevisti individuati, al fine di utilizzare anch'essi per ottimizzare le pratiche didattiche.

L'analisi *locale*, se richiesta dalle singole scuole, potrà restituire un confronto accurato dei comportamenti individuati sui dati di un certo istituto paragonandoli alle scuole dello stesso territorio o della stessa tipologia. Il trattamento anonimizzato dei dati non permetterà in alcun modo di estrarre informazioni né a livello di singolo studente, né a livello di singolo docente, né a livello di sezione. Il minimo livello di granularità previsto dalla ricerca è per classe e materia.

L'obiettivo non è certo quello di sostituirsi all'analisi della situazione scolastica effettuata da dirigenti scolastici e docenti che, in base alla loro esperienza e al contatto diretto con gli studenti, sono in grado di verificare in modo preciso e soggettivo l'andamento educativo e didattico, ma di fornire un ulteriore ausilio cercando di evidenziare informazioni che risultano essere nascoste nel grande volume dei dati e che possono o meno confermare le loro valutazioni.

L'aspetto scientifico del progetto è seguito da

- *KDD Lab - Knowledge Discovery and Data Mining Laboratory*
- *ISTI - Istituto di Scienza e Tecnologie dell'Informazione - CNR*
- *Dipartimento di Ingegneria e Architettura - Università di Parma*
- *Dipartimento di Informatica - Università di Pisa*

partner tecnici: *Gruppo Spaggiari, ARGO Software*
con il patrocinio dall'*Assemblea Legislativa della Regione Emilia-Romagna*

Riferimenti bibliografici

1. Romero, C., & Ventura, S. (2010). Educational data mining: a review of the state of the art. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 40(6), 601-618.
2. Márquez-Vera, C., Cano, A., Romero, C., & Ventura, S. (2013). Predicting student failure at school using genetic programming and different data mining approaches with high dimensional and imbalanced data. Applied intelligence, 38(3), 315-330.
3. Bishop, C. M. (2006). Pattern recognition. Machine Learning, 128, 1-58.
4. Engelbrecht, A. P. (2007). Computational intelligence: an introduction. John Wiley & Sons.
5. Ferguson, R. (2012). Learning analytics: drivers, developments and challenges. International Journal of Technology Enhanced Learning, 4(5-6), 304-317.
6. Codice di deontologia e di buona condotta per i trattamenti di dati personali per scopi statistici e scientifici, (Provvedimento del Garante n. 2 del 16 giugno 2004, Gazzetta Ufficiale 14 agosto 2004, n. 190)