Electronic Logbook Data Mining

Applicazione di tecniche di Educational Data Mining ai dati dei registri elettronici: approccio "data-driven" per l'analisi delle dinamiche didattiche

Giulio Angiani [1], Alberto Ferrari [1], Fosca Giannotti [3], Monica Mordonini [1]
Dino Pedreschi [2], Agostino Poggi [1], Elena Salvatori [2], Michele Tomaiuolo [1]

[1]*Dipartimento di Ingegneria dell'Informazione Università di Parma*
[2]*Dipartimento di Informatica Università di Pisa*
[3]*Istituto di Scienza e Tecnologie dell'Informazione - CNR*

Enti coinvolti

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

Con il Patrocinio dell'Assemblea legislativa
della Regione Emilia-Romagna

Educational Data Mining

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 da loro ottenuti nelle singole prove di verifica, dalla tipologia di queste prove alla loro collocazione temporale nell'anno scolastico, ecc.
Le attuali tecniche di gestione dei big data associate a metodologie di Machine Learning permettono di ottenere informazioni qualitative e quantitative (data mining) derivanti dai dati grezzi [1,2].
L'obiettivo del nostro progetto è quello di recuperare e analizzare 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.
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.

Il Progetto - Che studio vogliamo fare

Lo scopo del progetto di studio è la ricerca di comportamenti e fenomeni sia ricorrenti che anomali che si possono verificare nel corso dell'anno scolastico dal punto di vista didattico e organizzativo nelle diverse scuole che aderiranno alla nostra iniziativa.

La nostra analisi utilizzerà i dati anonimizzati provenienti dai registri elettronici delle scuole che aderiranno al progetto.

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 [3,9].

In breve, si tratta di "osservare" i dati grezzi per confrontarli con quelli aggregati, frutto di analisi statistiche sui dati relativi a prove ministeriali nazionali o internazionali, al fine di individuare conferme o scostamenti di valutazione utili per un possibile intervento mirato nelle attività di supporto al recupero di situazioni problematiche.

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 però, per avere valore scientifico reale, ha spesso bisogno di una grande mole di dati (approccio Big-Data) che è, in questo momento, uno degli argomenti più interessanti della ricerca scientifica informatica.

Il Dataset, i dati educativi di cui abbiamo bisogno

Per il progetto, abbiamo bisogno di costruire un Dataset che ci permetta di seguire passo-passo l'evoluzione della situazione a livello del singolo studente durante le attività scolastiche.

Il Dataset conterrà dati estratti dagli applicativi di gestione dei registri elettronici utilizzati a livello di singola scuola. Oltre alla classe frequentata, l'indirizzo di studio, la scuola di appartenenza (città, tipo di scuola), verranno registrati i dati relativi alla frequenza scolastica, alle valutazioni ricevute durante l'anno, ai voti di fine periodo (quadrimestre, fine anno, settembre).

Il Dataset non salverà alcuna informazione personale di studenti, docenti e personale scolastico e sarà gestito ai sensi delle vigenti normative sul Trattamento dei Dati Personali ai fine di ricerca e statistica [10].

Ambiti di ricerca

Lo studio proposto sarà orientato a vari aspetti che qui elenchiamo nel dettaglio.

- *Ricerca 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 e che possono fornire importanti informazioni spesso non intercettabili con analisi locali che interessano pochi alunni o anche una singola scuola.
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.

- *Visualizzazione di dati educativi*

Data la natura multidimensionale dei dati educativi, è necessario uno studio delle 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.

- *Relazione tra misurazioni della scuola e risultati INVALSI*

Una seconda fase dello studio prenderà 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.

- *Predire insuccessi a partire dalle misurazioni e dai risultati INVALSI*

A partire dai dati delle misurazioni di un anno scolastico e dai risultati dei test INVALSI (solo per le classi seconde) si cercheranno correlazioni (se esistono) con l'andamento degli studenti nell'anno scolastico successivo.
L'obiettivo, nel caso siano presenti legami evidenti, è di predire delle difficoltà con un preavviso di un anno. 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 [4].
Il nostro obiettivo è di utilizzare tecniche di analisi dei dati di tipo educativo nel contesto italiano attualizzandole alle metriche e alle modalità di valutazione presenti nella realtà della nostra scuola.

Presentazione dei risultati e restituzione alle scuole

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 far emergere i comportamenti e le pratiche migliori in modo da poter contaminare le altre scuole nell’ottica di un miglioramento continuo dell’offerta formativa.
Particolare attenzione sarà data nel comunicare 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 partner, potrà restituire un confronto accurato dei comportamenti individuati sui dati di un certo istituto paragonandoli alle scuole dello stesso territorio e/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/materia.

Bibliografia e letteratura sul tema

[1] Bishop, Christopher M. "Pattern recognition and *Machine Learning*". (2006). Springer-Verlag New York

[2] Engelbrecht, Andries P. "*Computational intelligence: an introduction"*. John Wiley & Sons, 2007.

[3] C. Romero, S. Ventura. Educational Data Mining: A Review of the State-of-the-Art. IEEE Transaction on Systems, Man, and Cybernetics, Part C: Applications and Reviews. 40(**6**), 601-618, 2010.

[4] C. Márquez-Vera A. Cano, C. Romero S. Ventura (2013). "Predicting student failure at school using genetic programming and different data mining approaches with high dimensional and imbalanced data". In: Applied Intelligence, Springer 38/3, pp. 315–330.

[5] OECD (2013a). "Better Skills, Better Jobs, Better Lives". In: OECD Publishing.

[6] OECD (2013b). "OECD Skills Outlook 2013: First Results from the Survey of Adult Skills". In: OECD PUblishing.

[7] OECD, "PISA 2012 results in Italy", In: OECD PUblishing.

[8] Le rilevazioni degli apprendimenti INVALSI, A.S. 2014-15, Rapporto di Sintesi, INVALSI, 2015.

[9] Ferguson, Rebecca (2012). "Learning analytics: drivers, developments and challenges". In: International Journal of Technology Enhanced Learning, pp. 304–317.

[10] 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)