

APPROCCIO CORRETTO ALL'ANALISI DEI DATI PER LE IMPRESE INDUSTRIALI



Negli ultimi tempi si parla moltissimo di dati, della loro analisi e di tutte le possibili implicazioni. Sui giornali si legge di

Big Data che condizionano la nostra esistenza.

Purtroppo spesso chi scrive non ha un'idea precisa dell'argomento ma semplicemente ripete frasi ad effetto.

L'analisi dei dati si basa in buona parte sulla statistica e in alcuni concetti non è certo di recente ideazione. La disponibilità diffusa di computer e di capacità di memoria a costi bassi ha permesso di attuare i concetti ipotizzati non realizzabili fino a poco tempo fa. **L'analisi dei dati si è evoluta successivamente nelle attuali forme di machine learning.**

Le aziende industriali e non solo quelle stanno iniziando a mettere a frutto i dati, trasformandoli in informazioni. Questo processo non è sempre così breve.



Nelle fabbriche si è partiti con i PLC (programmable logic controller), introducendo

progressivamente sistemi di controllo più integrati e complessi, come ideali prodromi del processo.

Le tecnologie permettono una approfondita comprensione dei processi tramite serie di dati con molte variabili e su scale temporali più o meno lunghe. Utilizzando mesi o persino anni di informazioni, magari sepolte negli archivi delle aziende, adattate in una forma comprensibile dagli algoritmi, queste possono essere integrate con i sistemi di controllo esistenti, es. PLC, e utilizzate per

ottimizzare la gestione globale.

I modelli analitici possono individuare soluzioni operative efficienti basandosi su parametri fisici controllabili, come velocità, vibrazioni, pressioni o correnti.

Nel processo, l'integrazione delle diverse competenze, meccaniche, elettriche e elettroniche dei tecnici è la chiave di successo per le aziende che cercano di sfruttare i dati. **L'applicazione dell'analisi e del machine learning è possibile solo dopo che i dati di processo sono stati analizzati, arricchiti.**

I passi principali per costruire un sistema di analisi in azienda

Parlando di aziende industriali, la prima operazione è quella di delineare le fasi del processo produttivo con esperti delle linee o degli impianti, segmentando



le diverse operazioni e trattamenti. **Si identificano i componenti o le operazioni chiave che determinano i risultati del processo stesso, quindi i parametri, i motori e i meccanismi critici.** Questi ultimi sono importanti sia per eventuali operazioni che possono fermare la produzione, sia a causa di manutenzioni onerose.

A questo punto occorre identificare i tipi di sensori e la loro allocazione, non

	Machine	Sensor_1	Sensor_2	Sensor_3	Sensor_4	Sensor_5	Sensor_6	Sensor_7
0	M_0001	2.633	0.918	4.229	13.792	23.403	13.764	6.054
1	M_0002	9.244	22.732	15.307	8.553	5.707	2.999	1.428
2	M_0003	3.183	28.526	8.735	4.898	3.685	2.960	2.398

Tabella dati in Pandas Python

dimenticando gli eventuali PLC, anche loro fonte preziosa di informazioni da integrare.

Occorre poi valutare le relazioni ovvero le equazioni che governano

il processo (come le caratteristiche delle macchine utensili o le relazioni termodinamiche), così come le altre possibili variabili coinvolte.

I tecnici delle linee produttive si debbono unire agli specialisti della analisi per validare la simulazione del processo. Si osservano quindi i primi dati provenienti dai sensori, insieme alle eventuali ridondanze dei sensori stessi e alle precisioni degli strumenti.

Gli addetti degli impianti hanno un ruolo fondamentale in questa fase.

A questo punto si inizia a lavorare davvero sui dati...

I dati raccolti dal processo contengono quasi sempre dei problemi e occorre sistamarli prima di procedere oltre.

I valori sono normalmente raccolti in forma di matrice per essere elaborati e possono giungere nelle forme più svariate. Quelle più comuni possono essere "csv" oppure "json" ma non solo.

Similarmente ad un foglio Excel, **la forma matriciale normalmente prevede ad esempio le osservazioni di una specifica macchina sulle righe e le caratteristiche dei vari sensori sulle colonne.**

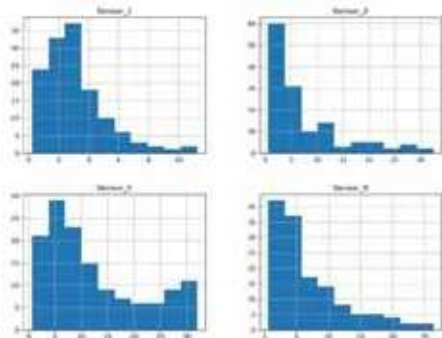
Questa è una delle fasi più lunghe e impegnative del lavoro e nel gergo si chiama "data wrangling".

Bisogna verificare i dati mancanti e considerarne l'eventuale sostituzione con un valore, spesso

medio o mediano, oppure farli indovinare al sistema con un piccolo artificio o, in ultima istanza, cancellare la serie a cui appartengono.

Può succedere di avere valori strani ma

può essere colpa di sensori difettosi o di errori di rilevazione: qui l'esperienza umana è insostituibile. A volte si giunge a combinare caratteristiche diverse fra loro per produrne di nuove.



Istogrammi distribuzione dati

Importante è comunque dare al numero un valore fisicamente corretto.

Durante l'elaborazione occorre fare attenzione, ad esempio, se siamo interessati a regimi stazionari oppure a transitori, per cui le nostre scelte potrebbero essere differenti.

Si studiano a questo punto le distribuzioni statistiche e le eventuali correlazioni fra i dati stessi, eliminando quelle che non danno valore ma aggiungono solo variabili al calcolo.

"The curse of dimensionality"

	Sensor_1	Sensor_2	Sensor_3	Sensor_4	Sensor_5	Sensor_6	Sensor_7
Sensor_1	1.000000	0.672929	0.345505	-0.142894	-0.392897	-0.210781	-0.218172
Sensor_2	0.672929	1.000000	0.361688	-0.346619	-0.524193	-0.382358	-0.304106
Sensor_3	0.345505	0.361688	1.000000	-0.217425	-0.647803	-0.471863	-0.374171
Sensor_4	-0.142894	-0.346619	-0.217425	1.000000	-0.159215	-0.125506	-0.120707
Sensor_5	-0.392897	-0.524193	-0.647803	-0.159215	1.000000	0.408091	0.262668
Sensor_6	-0.210781	-0.382358	-0.471863	-0.125506	0.408091	1.000000	0.744915
Sensor_7	-0.218172	-0.304106	-0.374171	-0.120707	0.262668	0.744915	1.000000

Tabella correlazione dati (Pearson)

La "maledizione della dimensionalità": con una quantità fissa di campioni numerici, il potere predittivo medio cresce all'aumentare del numero di dimensioni o di caratteristiche

utilizzate, ma oltre una certa dimensionalità inizia a deteriorarsi invece di migliorare costantemente (*da Wikipedia*). Generalmente **il numero delle osservazioni (righe) deve superare il numero di caratteristiche (colonne) per produrre un modello affidabile.** È comunque meglio ridurre per quanto possibile il numero delle caratteristiche stesse, per non appesantire troppo il calcolo.

Il contributo dei computer

I processi industriali possono essere caratterizzati da fenomeni deterministici e probabilistici. **Con l'aiuto dell'esperienza e della statistica, utilizzando algoritmi di varie tipologie, si giunge a creare modelli che rispecchiano le caratteristiche di funzionamento degli impianti.**

Da questi modelli possiamo ottenere incrementi di produttività, miglioramenti nei consumi di energia o indicazioni sulla cadenza della manutenzione.

Occorre provare diversi modelli, ai quali deve essere dato un riscontro fisico, quindi ci vuole normalmente anche un certo tempo.

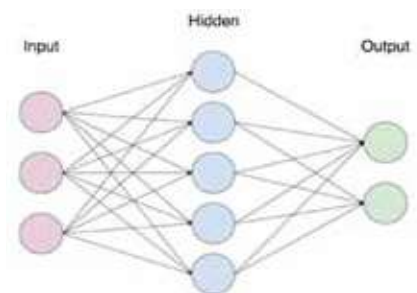
Prima di passare alle reti neurali, che sono di difficile interpretazione

del risultato, si possono utilizzare algoritmi più tradizionali ma egualmente molto efficienti.

La realizzazione dell'analisi dei dati nell'industria richiede gruppi inter-

funzionali composti da operatori, data scientist, ingegneri ed esperti di processo.

Le aziende devono migliorare le competenze negli strumenti di analisi; gli addetti hanno normalmente una formazione tecnica o ingegneristica e sono abituati a sfruttare le formule per descrivere i processi fisici.



Rete neurale

Acquisite queste competenze, gli esperti possono supportare meglio la collaborazione con i data scientist per affrontare i problemi e per selezionare i dati utili e i modelli corretti.

Le fasi quindi per l'applicazione industriale dell'analisi sono:

- Creazione di set di dati, importandoli, analizzandoli e ripulendoli.
- Addestramento di un modello, includendo la sperimentazione, i test di ipotesi e l'ottimizzazione dei parametri.
- Aggiornamento del modello e eventuale ricostruzione quando sono disponibili nuovi dati.
- Valutazione e confronto del modello con le versioni esistenti.
- Distribuzione ed utilizzo del modello per previsioni.

Nei prossimi articoli sull'argomento vedremo di approfondire i vari passi del processo.

Grafici e immagini dell'autore:

<http://www.agenda.unict.it>;

<https://pngimage.net>; <https://hackerstribе.com>