Failure prediction for batch jobs

Alessio Arcara

Base di dati messa a disposizione

Stato dei jobs in esecuzione

id	ts	jobid	idx	queue	hn	js	nc	hsj	hsm	cpt	rt	owner	rss	swp	sn	disk
3591906582	1643567948	4529008	0	belle	wn-204-13-31-08-a	2	1	10.0	400.0	2329	2775	belleprd	1.527156	4.098376	ce06-htc	2.643931
3591906583	1643567948	4529010	0	belle	wn-200-13-01-08-a	2	1	11.0625	354.0	2284	2775	belleprd	1.597328	4.14604	ce06-htc	2.64806
3591906584	1643567948	4529011	0	belle	wn-200-10-31-07-a	2	1	11.0625	354.0	2314	2775	belleprd	1.538676	4.080308	ce06-htc	2.668125
3591906585	1643567948	4529012	0	belle	wn-200-08-27-03-a	2	1	10.75	172.0	2295	2775	belleprd	1.548	4.2612	ce06-htc	2.653554
3591906586	1643567948	4529013	0	belle	wn-200-08-07-02-a	2	1	10.75	172.0	2286	2775	belleprd	1.531988	4.062452	ce06-htc	2.655122

Dati di accounting

	id	jobid	queue	fromhost	exechosts	starttimeepoch	eventtimeepoch	exitstatus	jobstatus	maxrmem	maxrswap	numprocessors
0	36471725	2873422	clas12vo	ce01-htc	wn-205-11-05-01-a	1620776362	1620779226	0	4	14936	1773164	1
1	36406413	3693170	atlas	ce03-htc	wn-200-10-11-12-a	1620738557	1620754559	0	4	1488360	5452924	1
2	85214769	5731930	lhcb	ce04-htc	wn-205-11-05-02-a	1648487169	1648487324	0	4	65888	1846256	1
3	85214770	5702700	belle	ce04-htc	wn-200-13-01-08-a	1648410179	1648487296	0	4	1828668	4489816	1
4	36471728	2873430	clas12vo	ce01-htc	cn-608-02-08	1620776516	1620779235	0	4	14312	1962080	1

Dallo stato dei jobs campionati **ogni 3 minuti** dal batch system e dai dati di accounting si sono estratti il consumo di **RAM**, **DISCO** e **SWAP** della prima ora di ogni job raggruppando per 15 minuti.

task di Machine Learning visti

- Job runtime prediction

  avere una stima del tempo di esecuzione di un job

- Job failure prediction

  predire se un job fallirà

- Job “zombie” prediction

  predire se un job entrerà in stato “inerte”

Job “zombie” prediction

Nel solo mese di settembre 2021 **5862 slots** delle macchine sono rimasti occupati per **41034 giorni** prima di essere eliminati dal batch system!

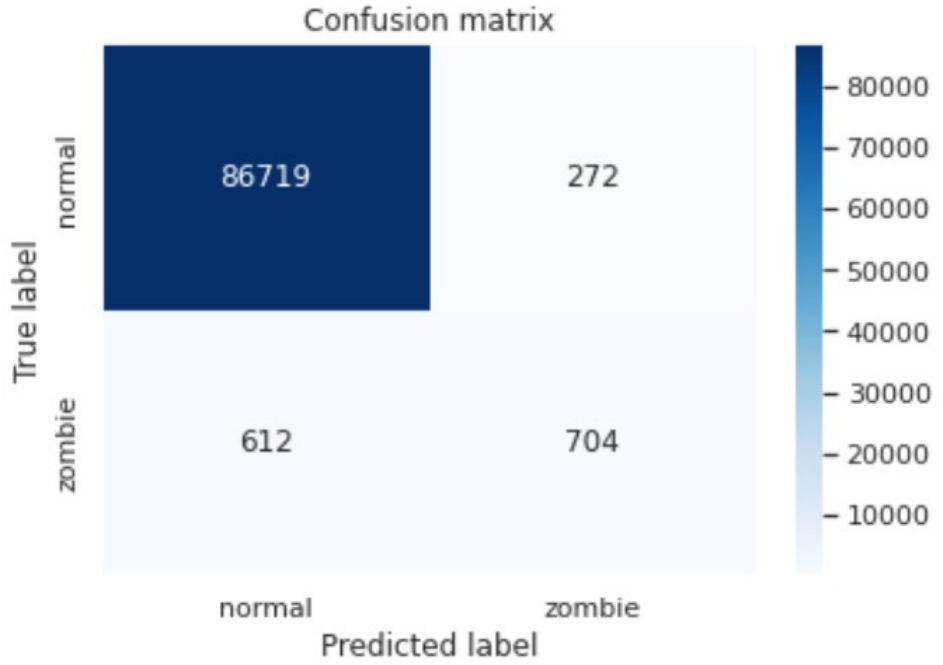
queue	too_much_time	size	perc	time_lost
atlas	5187	257036	2.018005	36309
alice	260	122209	0.212750	1820
lhcb	155	300598	0.051564	1085
cms	129	11690	1.103507	903
belle	77	79647	0.096677	539
clas12vo	31	6325	0.490119	217
muoncoll	11	86	12.790698	77
dampe	8	18209	0.043934	56
na62	3	9304	0.032244	21
ams	1	41679	0.002399	7

Stato attuale

- Sulla prima metà di settembre è stato addestrato XGBoost, un modello di Machine Learning.
- Si è ottenuto sulla seconda metà di settembre un 72% di **accuratezza** sulla classe meno rappresentata.
- Sulla base di questi risultati, è stato creato uno script per controllare le performance del modello con logs freschi.

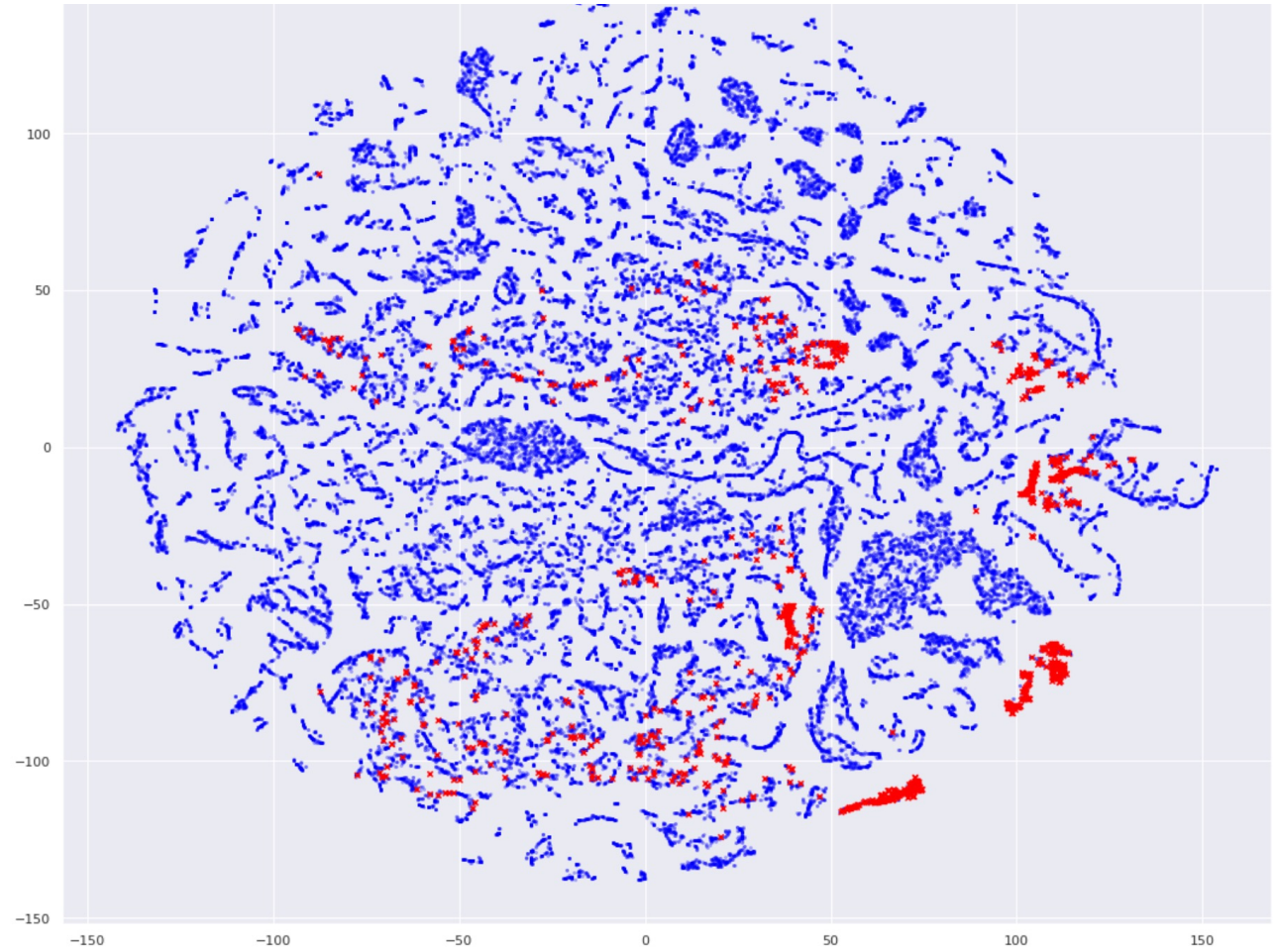
```
*** looking for logs at -> ../../test/ ***
*** getting the predictions for 956 unfinished jobs ***
*** saved the predictions to -> '/home/jovyan/notebooks/
```

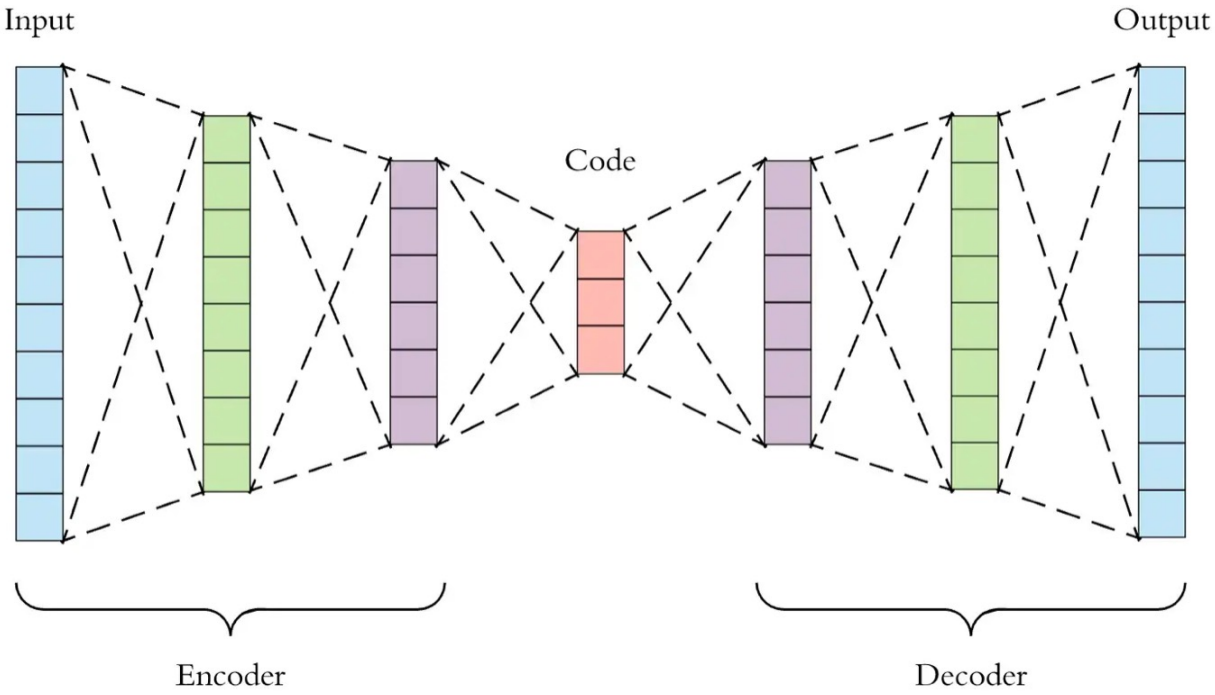
*** Confusion matrix ***



*** Precision, Recall, F1-measure per classe e media ***

	normal	zombie	all
precision	0.992992	0.721311	0.857152
recall	0.996873	0.534954	0.765914
f1_measure	0.994929	0.614311	0.804620





Rappresentazione 2D dello spazio latente di un autoencoder

Si nota che ALCUNI tipi di “zombie” sono identificabili con sicurezza: quelli dei **cluster rossi** ben definiti.

Prossimi passi

Le informazioni fornite al modello sono limitate, arricchirle con:

- L’**intera vita** del job piuttosto che solo la prima ora
- lo **stato delle macchine** incrociato con i jobs attualmente eseguiti su quelle macchine

Il numero di jobs normali e ‘zombie’ è **altamente sbilanciato**:

- Provare approcci di Anomaly Detection

Grazie per l’attenzione!