Progetto

transactional systems & data warehouse :

Simulazione ed Analisi di un data set di Impianti Ansaldo

# Introduzione

Circa un mese fa, intorno alla metà di Marzo del 2018, il sottoscritto, Mauro Angelini, Il Professor Roberto Sacile, l’Ingegner Alessandro Sorce e il collega di Ingegneria meccanica, stagista alla Ansaldo, Simone Ghettini ci incontravamo per la prima volta per fornire una prima struttura al progetto .

Il problema per cui siamo stati incaricati di cercare di ottenere dei risultati informativi può esser delineato nel seguente modo : abbiamo un impianto costituito da una centrale che scarica calore verso l’ambiente. Tale calore, sotto forma di vapore, viene raccolto una volta terminata la sua condensazione. Il carico di calore scaricato dall’impianto va a costituire la variabilità e proprio tal parametro influenza il buon funzionamento della macchina.

I dati che ci vengono forniti sono relativi a diverse variabili descrittrici dell’impianto in un caso estivo.

Il nostro scopo è allora quello di studiare tali dati per analizzare le relazioni delle variabili in gioco.

Il Professor Sacile ci propone di approcciare l’analisi di questo fenomeno “sfruttando” la potenza delle Reti Neurali.

# Procedura risolutiva adottata

In questa sezione cercherò di spiegare quanto più possibile i passi che vanno a costituire la procedura adottata per la risoluzione del problema.

Come detto in precedenza il problema viene approcciato tramite la tecnica di regressione non lineare della rete neurale. Prima di tutto io e il collega Angelini ne abbiamo studiato la teoria e, tramite l’utilizzo del sistema MATLAB e quindi del suo toolbox per le reti neurali, abbiamo scritto i codici relativi alla approssimazione delle variabili “vuoto atteso” e “incondensabili” relativi, rispettivamente, a quelle che erano le consegne del “modello 1” e “modello 2” forniteci dall’ Ingegner Sorce e allegate in un opportuno file Excel.

Il secondo passo è stato allora quello di provare ad alterare e/o modificare e quindi ridurre o ristrutturare il data set da fornire alla rete neurale. Questo al fine di far emergere quelle che sono le variabili o i valori selezionati da esse che più vanno ad influenzare la variabile di output stimata dalla rete.

Per il perseguimento di tale scopo abbiamo pensato a tre funzioni distinte :

1. Un campionatore
2. Un alteratore della moda di una variabile
3. Un clusterizzatore

La prima funzione, dopo aver preso in input il valore relativo all’intervallo di campionamento con il quale si vuole effettuare la riduzione del data set da fornie alla rete procede selezionando una riga della matrice (le cui colonne sono le variabili di input e la variabile di output) con frequenza corrispondente all’intervallo di campionamento specificato. (Per esempio : se l’intervallo è relativo al valore di 10, allora ogni 10 righe prendo la riga).

I dati selezionati vengono così salvati in una matrice ausiliaria che costituirà il data set ridotto e i dati scartati vengono a loro volta salvati in un ulteriore matrice che sarà invece utilizzata per effettuare il testing sul modello generato dalla rete.

La seconda funzione, come si intuisce dal nome, ha il particolare obiettivo di alterare la moda di una data variabile che le viene data in input. Di fatto tal funzione opera andando anzitutto a calcolare la suddetta moda della variabile, quindi si procede definendo un intorno di valori con il centro nell’intorno nella moda ed infine sostituendo quei valori con degli altri arbitrari valori non appartenti a quel’ intorno. La scelta che ho effettuato su quest’ultimi valori selezionabili arbitrariamente è ricaduta su un banale insieme randomico non appartenente a quel calcolato intorno.

La terza ed ultima funzione, la meno banale delle due, è pensata sulla base di quanto riportato dal capitolo relativo alla “data preprocessing” (in particolare alla pagina 36) delle dispense del corso di data mining che riporto in allegato.

Essa esegue una riduzione dei dati in modo da suddividere il data set in ‘N’ partizioni (con ‘N’ numero delle partizioni selezionate dall’utente) e, di esse, solo un sottoinsieme verrà allora selezionato e fornito alla rete neurale. Ma come effettua la suddivisione? In modo omogeneo, il che significa che cercherà di andare a comporre gli insiemi in modo che le rispettive proprietà statistiche siano, tra le varie partizioni, il più equivalenti possibile. Per far ciò la funzione ricorre ad un’ulteriore funzione per il calcolo della distanza euclidea tra una data riga, della matrice del data set da ridurre, e tutte le altre; quindi lanciata quest’ultima funzione dalla riga di partenza, si procede andando a prelevare la riga la cui distanza è la minima tra tutte quelle calcolate dalla riga di partenza. Fatto ciò si và a metter tal riga in un nuovo insieme. La riga di partenza diventa allora questa seconda riga e si itera il procedimento.

Come si intuisce, due righe che hanno valori quasi equivalenti non potranno mai stare in uno stesso insieme. Ed inoltre, se N è abbastanza elevato, si riuscirà a costruire insiemi dove le righe che hanno valori quasi equivalenti non potranno mai trovarsi nello stesso insieme e, così facendo “livellerò” il numero delle righe ad alte frequenze. (Per esempio in un data set costituito da una variabile x di 10 righe, se il valore 8 è uguale a tre righe della matrice x, e ho scelto di partizionare in N=3 insiemi, per forza di cose le tre x(i) = 8 andranno in insiemi diversi e mai in uno singolo. Ecco che la frequenza di x(i)=8 se considerata in uno dei tre insiemi sarà ora uguale a uno.)

# Commento risultati

Tra i file in allegato al progetto è possibile trovare il documento Excel relativo alla tabella dei risultati ottunuti in relazione alla cinquantina di simulazioni/ esperimenti effettuati.

In questa sezione si cerca di dare una spiegazione, il più possibile completa, sulle relazioni dei vari esperimenti.

Anticipo che vengono valutati i seguenti casi :

1. La differenza tra i risultati ottenuti in relazione alla normalizzazione standard (quindi con dati normalizzati tra i valori 0 e 1) e la normalizzazione significativa (generante dati normalizzati sempre tra 0 e 1 ma usando non più i minimi e i massimi reali bensì dei minimi e massimi che sono stati ritenuti significativi dal collega Ghettini). Tale analisi viene effettuata su tre intervalli di campionamento distinti : ogni 100, ogni 10 e ogni uno, quindi non campionando ma prendendo il data set per intero. E, per ognuno di questi tre, vengono considerati una volta i casi di diversa clusterizzazione (quindi con numero di clusters differenti) mischiando il data set e una volta non mischiando. Per ognuno degli intervalli di campionamento considerati eseguo una clusterizzazione ovviamente distinta. (Dico ovviamente perché se per esempio campiono ogni 100 su un data set di partenza di 132944 dati terminerò in un data set di 1329, mentre nel caso di campionamento ogni 10 termino in un data set di 13294 dati. Ecco che di conseguenza scelgo di suddividere tali data set in un numero di insieme logicamente differente tra un caso e l’altro).

Nella fattispecie : nel caso 100 considero i casi :

1. Numero di cluser : 0 , di cui ne seleziono : 0
2. Numero di cluser : 25 , di cui ne seleziono : 5
3. Numero di cluser : 1000 , di cui ne seleziono : 200

Nel caso in cui campiono ogni 10 righe considero invece :

1. Numero di cluser : 0 , di cui ne seleziono : 0
2. Numero di cluser : 1000 , di cui ne seleziono : 200
3. Numero di cluser : 2000, di cui ne seleziono : 400
4. Numero di cluster: 10000, di cui ne seleziono: 2000

E infine, nel caso in cui non campiono considero :

1. Numero di cluser : 0 , di cui ne seleziono : 0
2. Numero di cluser : 1000, di cui ne seleziono : 200
3. Numero di cluser : 10000, di cui ne seleziono : 2000
4. I casi di cui sopra ho appena parlato vengono poi rianalizzati ma questa volta in relazione alla differenza tra gli esperimenti in cui vengono mischiati i dati e quelli in cui non mischio ma mantengo l’ordine temporale con cui sono stati raccolti. Questa volta ovviamente andrò a tener fisso in un caso la normalizzazione al caso standard e nell’altro la normalizzazione al caso significativo.
5. Prendo in esame un caso che credo sia interessante. Vado a confrontare i risultati ottenuti, tenendo fissa la normalizzazione a standard, in relazione ad un esperimento in cui campiono ogni 1000 righe ma non cluterizzo (quindi dando alla rete un data set per il training di circa 90 dati) e ad un secondo esperimento in cui invece non campiono ma clusterizzo il data set in 1000 cluster e prendendo, tra i 1000 cluster ottenuti, solo un cluster da fornire alla rete (ancora una volta quindi dando alla rete un data set di 90 dati). Sia chiaro che ogni volta considero sia il caso in cui mischio i dati e sia quello in cui mantengo l’ordine.

Il tutto allo scopo di valutare l’efficacia dell’algoritmo di clusterizzazione utilizzato.

1. Considero i quattro modelli forniti dall’Ingegner Sorce utilizzando il 90% dei dati totali per il testing.
2. Infine andiamo ad applicare la funzione di alterazione delle frequenze in un intorno della moda della variabile “ temperatura ambiente ” sul modello 1. L’esperimento va a modificare il valore di circa 22 mila valori in un range con centro nella suddetta moda +/- la soglia.

Eseguo i casi di cui sopra riportando i rispettivi valori dalla tabella excel:

* 1. Ipotizzo di non mischiare i dati ( Mixer = 0, riportando prima la normalizzazione standard e poi quella significativa):

a) Numero di cluser : 0 , di cui ne seleziono : 0

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Cluster (PF3)** | **Cluster\_IN (PF3)** | **NN\_Error** | | | **Data Test\_Err** | **Accuracy** | **Train Set Size** |
| **Train** | **Validation** | **Test** |
| 0 | 0 | 1.135E-03 | 1.248E-03 | 1.272E-03 | 1.132E-03 | 0.911 | 930 |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Cluster (PF3)** | **Cluster\_IN (PF3)** | **NN\_Error** | | | **Data Test\_Err** | **Accuracy** | **Train Set Size** |
| **Train** | **Validation** | **Test** |
| 0 | 0 | 1.615E-03 | 1.489E-03 | 1.548E-03 | 1.635E-03 | 0.962 | 930 |

b) Numero di cluser : 25 , di cui ne seleziono : 5

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 25 | 5 | 1.830E-03 | 1.212E-03 | 1.185E-03 | 1.654E-03 | 0.986 | 188 |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 25 | 5 | 1.505E-03 | 2.428E-03 | 1.696E-03 | 2.478E-03 | 0.948 | 188 |

c) Numero di cluser : 1000 , di cui ne seleziono : 200

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 1000 | 200 | 5.196E-04 | 2.867E-04 | 3.862E-04 | 1.164E-02 | 0.724 | 280 |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 1000 | 200 | 1.583E-03 | 1.972E-03 | 2.210E-03 | 7.415E-03 | 0.320 | 280 |

Morale: la precisione è migliore nello standard in tutti i casi a parte a) (in quanto, probabilmente, il Train Set Size è maggiore)

d) Numero di cluser : 0 , di cui ne seleziono : 0

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 0 | 0 | 9.739E-04 | 1.029E-03 | 9.515E-04 | 9.755E-04 | 0.948 | 9307 |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 0 | 0 | 1.528E-03 | 1.692E-03 | 1.485E-03 | 1.541E-03 | 0.968 | 9307 |

e) Numero di cluser : 1000 , di cui ne seleziono : 200

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 1000 | 200 | 4.732E-04 | 4.727E-04 | 5.173E-04 | 1.598E-03 | 0.883 | 1960 |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 1000 | 200 | 9.330E-04 | 8.836E-04 | 9.287E-04 | 6.130E-03 | 0.855 | 1960 |

f) Numero di cluser : 2000, di cui ne seleziono : 400

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 2000 | 400 | 7.789E-04 | 7.918E-04 | 7.046E-04 | 3.336E-02 | 0.652 | 1960 |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 2000 | 400 | 1.389E-03 | 9.486E-04 | 1.106E-03 | 7.431E-03 | 0.815 | 1960 |

g) Numero di cluster: 10000, di cui ne seleziono: 2000

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 10000 | 2000 | 5.046E-04 | 4.989E-04 | 4.958E-04 | 5.490E-03 | 0.742 | 2800 |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 10000 | 2000 | 1.085E-03 | 9.939E-04 | 9.784E-04 | 1.643E-02 | 0.731 | 2800 |

Morale: aumentando il numero di dati forniti alla rete, l’errore sul testing è peggiore nel caso di normalizzazione significativa.

h) Numero di cluser : 0 , di cui ne seleziono : 0

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 0 | 0 | 9.948E-04 | 1.008E-03 | 9.955E-04 |  | 0.944 | 93060 |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 0 | 0 | 1.590E-03 | 1.608E-03 | 1.614E-03 |  | 0.965 | 93060 |

1. Numero di cluser : 1000, di cui ne seleziono : 200

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 1000 | 200 | 7.332E-04 | 6.845E-04 | 6.974E-04 | 1.462E-03 | 0.915 | 18620 |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 1000 | 200 | 1.314E-03 | 1.380E-03 | 1.373E-03 | 1.992E-03 | 0.952 | 18620 |

j) Numero di cluser : 10000, di cui ne seleziono : 2000

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 10000 | 2000 | 5.567E-04 | 5.326E-04 | 5.419E-04 | 3.327E-03 | 0.810 | 19600 |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 10000 | 2000 | 9.691E-04 | 9.565E-04 | 1.041E-03 | 5.304E-03 | 0.855 | 19600 |

Morale: la precisione è sempre peggiore nel caso di normalizzazione standard, mentre gli errori di trainig validation e testing risultano sempre maggiori nel caso della normalizzazione significativa.

* 1. Ipotizzo di non mischiare i dati ( Mixer = 1, riportando prima la normalizzazione standard e poi quella significativa):

a) Numero di cluser : 0 , di cui ne seleziono : 0

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 0 | 0 | 1.024E-03 | 1.217E-03 | 1.057E-03 | 1.179E-03 | 0.923 | 930 |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 0 | 0 | 1.340E-03 | 1.681E-03 | 1.573E-03 | 1.621E-03 | 0.968 | 930 |

b) Numero di cluser : 25 , di cui ne seleziono : 5

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 25 | 5 | 1.396E-03 | 1.807E-03 | 3.010E-03 | 2.488E-03 | 0.877 | 188 |
| 25 | 5 | 1.091E-03 | 1.438E-03 | 2.081E-03 | 4.000E-03 | 0.911 | 188 |

c) Numero di cluser : 1000 , di cui ne seleziono : 200

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 1000 | 200 | 5.071E-04 | 5.928E-04 | 7.404E-04 | 3.141E-02 | 0.668 | 280 |
| 1000 | 200 | 2.332E-03 | 3.091E-03 | 2.064E-03 | 1.696E-03 | 0.869 | 280 |

Morale: ?

d) Numero di cluser : 0 , di cui ne seleziono : 0

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 0 | 0 | 9.840E-04 | 1.009E-03 | 1.091E-03 | 1.006E-03 | 0.942 | 9307 |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 0 | 0 | 1.567E-03 | 1.462E-03 | 1.696E-03 | 1.576E-03 | 0.966 | 9307 |

e) Numero di cluser : 1000 , di cui ne seleziono : 200

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 1000 | 200 | 6.411E-04 | 5.906E-04 | 7.431E-04 | 6.882E-03 | 0.755 | 1960 |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 1000 | 200 | 7.845E-04 | 8.348E-04 | 7.543E-04 | 2.728E-03 | 0.912 | 1960 |

f) Numero di cluser : 2000, di cui ne seleziono : 400

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 2000 | 400 | 3.897E-04 | 3.143E-04 | 5.054E-04 | 6.625E-03 | 0.669 | 1960 |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 2000 | 400 | 8.811E-04 | 8.235E-04 | 9.552E-03 | 2.575E-03 | 0.993 | 1960 |

g) Numero di cluster: 10000, di cui ne seleziono: 2000

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 10000 | 2000 | 7.364E-04 | 7.035E-04 | 7.079E-04 | 2.001E-02 | 0.622 | 2800 |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 10000 | 2000 | 1.733E-03 | 2.044E-03 | 1.664E-03 | 9.756E-03 | 0.799 | 2800 |

Morale: aumentando il numero di dati forniti alla rete, l’errore sul testing è peggiore nel caso di normalizzazione significativa. La precisione è sempre peggiore in standard.

h) Numero di cluser : 0 , di cui ne seleziono : 0

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 0 | 0 | 9.936E-04 | 9.935E-04 | 9.881E-04 |  | 0.946 | 93060 |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 0 | 0 | 1.566E-03 | 1.587E-03 | 1.558E-03 |  | 0.967 | 93060 |

1. Numero di cluser : 1000, di cui ne seleziono : 200

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 1000 | 200 | 9.813E-04 | 9.912E-04 | 1.014E-03 | 1.304E-03 | 0.918 | 18620 |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 1000 | 200 | 1.277E-03 | 1.309E-03 | 1.310E-03 | 1.952E-03 | 0.956 | 18620 |

j) Numero di cluser : 10000, di cui ne seleziono : 2000

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 10000 | 2000 | 5.943E-04 | 5.981E-04 | 5.961E-04 | 4.593E-03 | 0.774 | 19600 |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 10000 | 2000 | 9.697E-04 | 9.700E-04 | 9.331E-04 | 5.902E-03 | 0.851 | 19600 |

Morale: la precisione è sempre peggiore nel caso di normalizzazione standard, mentre gli errori di trainig validation e testing risultano sempre maggiori nel caso della normalizzazione significativa.

2.1 Ipotizzo di non mischiare i dati ( Mixer = 0, riportando prima la normalizzazione standard e poi quella significativa):

a) Numero di cluser : 0 , di cui ne seleziono : 0

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Cluster (PF3)** | **Cluster\_IN (PF3)** | **NN\_Error** | | | **Data Test\_Err** | **Accuracy** | **Train Set Size** |
| **Train** | **Validation** | **Test** |
| 0 | 0 | 1.135E-03 | 1.248E-03 | 1.272E-03 | 1.132E-03 | 0.911 | 930 |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 0 | 0 | 1.024E-03 | 1.217E-03 | 1.057E-03 | 1.179E-03 | 0.923 | 930 |

b) Numero di cluser : 25 , di cui ne seleziono : 5

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 25 | 5 | 1.830E-03 | 1.212E-03 | 1.185E-03 | 1.654E-03 | 0.986 | 188 |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 25 | 5 | 1.396E-03 | 1.807E-03 | 3.010E-03 | 2.488E-03 | 0.877 | 188 |

c) Numero di cluser : 1000 , di cui ne seleziono : 200

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 1000 | 200 | 5.196E-04 | 2.867E-04 | 3.862E-04 | 1.164E-02 | 0.724 | 280 |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 1000 | 200 | 5.071E-04 | 5.928E-04 | 7.404E-04 | 3.141E-02 | 0.668 | 280 |

Morale: in generale l’errore sul test è peggiore nel caso in cui mischio

d) Numero di cluser : 0 , di cui ne seleziono : 0

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 0 | 0 | 9.739E-04 | 1.029E-03 | 9.515E-04 | 9.755E-04 | 0.948 | 9307 |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 0 | 0 | 9.840E-04 | 1.009E-03 | 1.091E-03 | 1.006E-03 | 0.942 | 9307 |

e) Numero di cluser : 1000 , di cui ne seleziono : 200

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 1000 | 200 | 4.732E-04 | 4.727E-04 | 5.173E-04 | 1.598E-03 | 0.883 | 1960 |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 1000 | 200 | 6.411E-04 | 5.906E-04 | 7.431E-04 | 6.882E-03 | 0.755 | 1960 |

f) Numero di cluser : 2000, di cui ne seleziono : 400

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 2000 | 400 | 7.789E-04 | 7.918E-04 | 7.046E-04 | 3.336E-02 | 0.652 | 1960 |
| 2000 | 400 | 3.897E-04 | 3.143E-04 | 5.054E-04 | 6.625E-03 | 0.669 | 1960 |

g) Numero di cluster: 10000, di cui ne seleziono: 2000

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 10000 | 2000 | 5.046E-04 | 4.989E-04 | 4.958E-04 | 5.490E-03 | 0.742 | 2800 |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 10000 | 2000 | 7.364E-04 | 7.035E-04 | 7.079E-04 | 2.001E-02 | 0.622 | 2800 |

Morale: non sembrano esserci signiicative variazioni all’aumentare del numero dei valori.

h) Numero di cluser : 0 , di cui ne seleziono : 0

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 0 | 0 | 9.948E-04 | 1.008E-03 | 9.955E-04 |  | 0.944 | 93060 |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 0 | 0 | 9.936E-04 | 9.935E-04 | 9.881E-04 |  | 0.946 | 93060 |

1. Numero di cluser : 1000, di cui ne seleziono : 200

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 1000 | 200 | 7.332E-04 | 6.845E-04 | 6.974E-04 | 1.462E-03 | 0.915 | 18620 |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 1000 | 200 | 9.813E-04 | 9.912E-04 | 1.014E-03 | 1.304E-03 | 0.918 | 18620 |

j) Numero di cluser : 10000, di cui ne seleziono : 2000

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 10000 | 2000 | 5.567E-04 | 5.326E-04 | 5.419E-04 | 3.327E-03 | 0.810 | 19600 |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 10000 | 2000 | 5.943E-04 | 5.981E-04 | 5.961E-04 | 4.593E-03 | 0.774 | 19600 |

Morale: all’aumenta dei valori, come noto dalla statistica i valori tendono a essere equivalenti.

* 1. Ipotizzo di non mischiare i dati ( Normalizzazione = STD, riportando prima il caso non mischiato e poi il caso mischiato ):

a) Numero di cluser : 0 , di cui ne seleziono : 0

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Cluster (PF3)** | **Cluster\_IN (PF3)** | **NN\_Error** | | | **Data Test\_Err** | **Accuracy** | **Train Set Size** |
| **Train** | **Validation** | **Test** |
| 0 | 0 | 1.615E-03 | 1.489E-03 | 1.548E-03 | 1.635E-03 | 0.962 | 930 |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 0 | 0 | 1.340E-03 | 1.681E-03 | 1.573E-03 | 1.621E-03 | 0.968 | 930 |

b) Numero di cluser : 25 , di cui ne seleziono : 5

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 25 | 5 | 1.505E-03 | 2.428E-03 | 1.696E-03 | 2.478E-03 | 0.948 | 188 |
| 25 | 5 | 1.091E-03 | 1.438E-03 | 2.081E-03 | 4.000E-03 | 0.911 | 188 |

c) Numero di cluser : 1000 , di cui ne seleziono : 200

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 1000 | 200 | 5.196E-04 | 2.867E-04 | 3.862E-04 | 1.164E-02 | 0.724 | 280 |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 1000 | 200 | 2.332E-03 | 3.091E-03 | 2.064E-03 | 1.696E-03 | 0.869 | 280 |

Morale : ?

d) Numero di cluser : 0 , di cui ne seleziono : 0

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 0 | 0 | 1.528E-03 | 1.692E-03 | 1.485E-03 | 1.541E-03 | 0.968 | 9307 |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 0 | 0 | 1.567E-03 | 1.462E-03 | 1.696E-03 | 1.576E-03 | 0.966 | 9307 |

e) Numero di cluser : 1000 , di cui ne seleziono : 200

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 1000 | 200 | 9.330E-04 | 8.836E-04 | 9.287E-04 | 6.130E-03 | 0.855 | 1960 |
| 1000 | 200 | 7.845E-04 | 8.348E-04 | 7.543E-04 | 2.728E-03 | 0.912 | 1960 |

f) Numero di cluser : 2000, di cui ne seleziono : 400

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 2000 | 400 | 1.389E-03 | 9.486E-04 | 1.106E-03 | 7.431E-03 | 0.815 | 1960 |
| 2000 | 400 | 8.811E-04 | 8.235E-04 | 9.552E-04 | 2.575E-03 | 0.993 | 1960 |

g) Numero di cluster: 10000, di cui ne seleziono: 2000

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 10000 | 2000 | 1.085E-03 | 9.939E-04 | 9.784E-04 | 1.643E-02 | 0.731 | 2800 |
| 10000 | 2000 | 1.733E-03 | 2.044E-03 | 1.664E-03 | 9.756E-03 | 0.799 | 2800 |

Morale: in generale all’aumentare dei dati di training l’errore è peggiore non mischiando.

h) Numero di cluser : 0 , di cui ne seleziono : 0

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 0 | 0 | 1.590E-03 | 1.608E-03 | 1.614E-03 |  | 0.965 | 93060 |
| 0 | 0 | 1.566E-03 | 1.587E-03 | 1.558E-03 |  | 0.967 | 93060 |

1. Numero di cluser : 1000, di cui ne seleziono : 200

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 1000 | 200 | 1.314E-03 | 1.380E-03 | 1.373E-03 | 1.992E-03 | 0.952 | 18620 |
| 1000 | 200 | 1.277E-03 | 1.309E-03 | 1.310E-03 | 1.952E-03 | 0.956 | 18620 |

j) Numero di cluser : 10000, di cui ne seleziono : 2000

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 10000 | 2000 | 9.691E-04 | 9.565E-04 | 1.041E-03 | 5.304E-03 | 0.855 | 19600 |
| 10000 | 2000 | 9.697E-04 | 9.700E-04 | 9.331E-04 | 5.902E-03 | 0.851 | 19600 |

Morale: la precisione è praticamente equivalente.

3.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 1000 | 1 | 1.274E-03 | 8.303E-04 | 1.296E-03 | 1.684E-03 | 0.889 | 93 |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 0 | 0 | 8.847E-04 | 1.992E-03 | 2.595E-03 | 1.621E-03 | 0.879 | 93 |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 1000 | 1 | 1.365E-03 | 1.215E-03 | 9.957E-04 | 1.549E-03 | 0.881 | 93 |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 0 | 0 | 3.549E-03 | 4.712E-03 | 1.080E-02 | 7.808E-03 | 0.510 | 93 |

Morale: nonostante il data set fornito alla rete sia in dimensione uguale in tutti e quattro i casi, sin nota una precisione del 50% nel caso in cui mischio ma non eseguo il clustering. Mentre eseguendo il clustering la precisione è dell’88%. Non mischiando la precisione si conferma sulle percentuali sopra l’85%.

Ed inoltre l’errore sul testing assume il valore minimo proprio nel caso di clustering in 1000 insiemi e prendendone solo uno.



|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **NN\_Error** | | | **Data Test\_Err** | **Train Set Size** |
| **Train** | **Validation** | **Test** |
| **M\_2a** | 7.974E-10 | 1.036E-09 | 1.561E-09 | 1.005E-09 | 9307 |
| **M\_2b** | 1.830E-09 | 2.393E-09 | 1.880E-09 | 1.974E-09 | 9307 |
| **M\_2c** | 1.184E-10 | 1.341E-10 | 1.458E-10 | 1.334E-10 | 9307 |
| **M\_1** | 9.843E-04 | 8.872E-04 | 1.032E-03 | 9.895E-04 | 9307 |

Morale: si nota immediatamente come l’errore minimo sul testing sia in corrispondenza del modello 2C (essendo che, come noto, le reti neurali, modellano più accuratamente se addestrate su un insieme di features elevato). E di conseguenza si osserva errore massimo nel caso del modello 1.

5.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Mixer (PF0)** | **Normalization** | **Tau (PF1)** | **Cluster\_IN (PF3)** | **P\_IN (PF3)** | **NN\_Error** | | | **Accuracy** | **Train Set Size** |
| **Train** | **Validation** | **Test** |
| NO | STD | 1 | 0 | 0 | 5.236E-03 | 5.276E-03 | 5.204E-03 | 0.687 | 93060 |
| 0 | 0 | 9.948E-04 | 1.008E-03 | 9.955E-04 | 0.948 | 93060 |

Morale: nella prima riga, dove viene riportato l’esperimento con l’alterazione della moda, si nota un abbassamento della precisione e un netto peggioramento nell’errore di testing. A dimostrazione di quanto il modello 1 sia influenzato da tal variabile.