THESIS PRESENTATION

Speech

- Presentazione - Slide 1

Buongiorno a tutti, sono Lai Nicolò. In questa breve discussione presento un'applicazione di reti neurali al data quality monitoring.

- Introduzione - Slide 2

L'utilizzo di algoritmi di deep learning porta un sostanziale aumento di flessibilità rispetto alle tecniche tradizionali. Inoltre, questi algoritmi risultano essere notevolmente accurati e si prestano in modo particolare all'automatizzazione delle procedure in cui sono coinvolti. Tutte queste caratteristiche risultano essere cruciali per un'applicazione al monitoraggio della qualità dei dati.

- Il Rivelatore - Slide 3

Per studiare la nuova procedura di data quality monitoring abbiamo sfruttato le camere a muoni presenti ai laboratori nazionali di Legnaro. Queste 4 camere a deriva si compongono di 4 layer da 16 tubi ciascuna.

Una specifica configurazione di elettrodi garantisce un campo elettrico approssimativamente uniforme all'interno dei tubi. Dunque, al passaggio di un muone, gli elettroni di ionizzazione derivano verso il filo anodico localizzato al centro della cella con una velocità pressochè costante. Questo permette di legare linearmente la posizione di passaggio del muone al tempo che gli elettroni impiegano a raggiungere il filo, ovvero il tempo di deriva.

Il rivelatore acquisce i dati con un sistema trigger-less. Ovvero, ogni volta che sul filo si deposita una carica superiore ad un threshold pre-impostato, il segnale viene inviato ad una scheda FPGA che implementa la time-to-digital conversion associando al segnale un time stamp t.

Per rivelare quindi il passaggio di un muone, vengono sfruttati due scintillatori che letti in coincidenza costituiscono un sistema di trigger esterno.

I due scintillatori forniscono inoltre un riferimento temporale t_0 relativo al passaggio del muone attraverso il detector: questo piedistallo temporale viene quindi utilizzato per il calcolo dei tempi di deriva degli elettroni liberati nel processo di ionizzazione.

- La Distributione dei Tempi - Slide 4

Abbiamo scelto di testare l'algoritmo di monitoring sulla distribuzione dei tempi di deriva, anche chiamata scatola dei tempi, includendo i contributi di tutti i tubi presenti nel detector. In un lavoro successivo a questa tesi, invece, stiamo considerando la distribuzione di ciascun singolo filo separatamente, in modo da aumentare la sensibilità della procedura e ricavarne più informazioni.

Come si può notare in figura, la distribuzione è approssimativamente uniforme con una larghezza di circa 400 ns. L'evidente picco sulla spalla sinistra è legato alla geometria del rivelatore, mentre i picchi presenti al centro della distribuzione sono solamente un artefatto del binning.

E' opportuno evidenziare che la scelta di monitorare la distribuzione dei tempi di deriva non è casuale: la forma e la larghezza di questa distribuzione infatti è strettamente legata ad una serie di possibili problematiche del detector. In altre parole, nel caso il rivelatore abbia un qualche tipo di malfunzionamento, come un elettrodo staccato, una fuga di gas dai tubi o una sbagliata tensione degli elettrodi, è altamente probabile che la distribuzione presenti delle anomalie.

- Overview dell'Algoritmo - Slide 5

L'algoritmo che abbiamo implementato per monitorare la qualità della distribuzione dei tempi di deriva può essere schematizzato come in figura.

Questo algoritmo si è mostrato essere estremamente flessibile. L'algoritmo nasce infatti per una ricerca model-independent di nuova fisica a LHC. Il fatto che sia la ricerca di nuova fisica e sia il data quality monitoring, seppur con obiettivi differenti, ricerchino delle discrepanze nei dati rispetto ad un modello di riferimento permette l'utilizzo di questo stesso algoritmo.

L'algoritmo richiede in input due datasets: un campione di riferimento R e un data sample D. Attraverso una rete neurale, quindi, il data sample D viene confrontato con il reference R implementando un test d'ipotesi. La forma della loss function della rete, infatti, riproduce il likelihood-ratio test con ipotesi nulla data dal reference e ipotesi alternativa parametrizzata dalla rete.

L'output del processo di training è dunque il test stastistic t, proporzionale al valore della loss function al termine del training, ed il maximum-likelihood fit del rapporto tra le due distribuzioni.

In particolare, il fit del rapporto delle due distribuzioni viene utilizzato per la localizzazione dell'anomalia nello spazio delle fasi, qualora essa sia presente. Il test statistic, invece, fornisce una stima della probabilità che una discrepanza sia presente nel data sample D. Per quantificare questa probabilità è necessario conoscere la distribuzione del test statistic: il teorema di Wilks ci assicura che sotto ipotesi nulla questa distribuzione tende asintoticamente ad un χ^2 con gradi di libertà pari, in questo caso, al numero di parametri liberi della rete neurale.

In pratica però, ciò non è sempre verificato. Per eliminare delle pericolose divergenze della loss function, infatti, abbiamo posto un vincolo superiore al valore assoluto dei parametri della rete, chiamto weight clipping. Questo parametro deve essere in primo luogo ottimizzato per garantire che la distribuzione del test statistic sotto l'ipotesi nulla si distribuisca effettivamente seguendo un χ^2.

Il primo step della procedura di monitoring è dunque l'ottimizzazione del weight clipping.

- Tuning della Rete - Slide 6

Abbiamo preso una distribuzione di riferimento, la 52, che è stata tenuta costante nel corso dei seguenti test. Il campione di riferimento viene costruito campionando 200000 tempi di deriva da questa run.

Per assicurarci di essere sotto l'ipotesi nulla, ovvero che il data sample D sia in accordo con il reference, vengono campionati diversi datasets con statistica inferiore sempre dalla stessa run. Viene fatto quindi girare l'algoritmo un numero di volte pari al numero di data sample campionati e viene costruita la distribuzione del test statistic in validità dell'ipotesi nulla.

- Tuning della Rete - Slide 7

Il weight clipping viene quindi scelto in modo da assicurare che la distribuzione dei test statistic sia in accordo con la distribuzione del χ^2 attesa.

Una volta trovato il weight clipping ottimale, questo può essere lasciato invariato, a patto di non alterare gli altri iperparametri della rete, come in particolare il rapporto tra la dimensione del reference R e del data sample D.

- Distribuzioni dei Tempi - Slide 8

In questa slide sono presenti le quattro distribuzioni dei tempi di deriva che abbiamo utilizzato per testare le performance dell'algoritmo. La prima in alto a sinistra, cioè la 52, è la distribuzione di riferimento, ovvero una distribuzione che gli esperti valutano come priva di anomalie.

La seconda, la 53, è estremamente simile a quella di riferimento, ed è stata utilizzata per assicurarci dell'assenza di errori di tipo 1, ovvero falsi positivi.

Le ultime due in basso, invece, sono state utilizzate per verificare la corretta rivelazione di anomalie da parte dell'algoritmo. In particolare, la 65 presenta una discrepanza sulla coda destra della distribuzione, mentre la 42 è caratterizzata da un evidente rumore aldifuori della scatola dei tempi.

- Test delle Performance - Slide 9

Dal processo di training della rete quindi ricaviamo il test statistic per ciascuna delle ultime tre distribuzioni confrontate con il reference: nel caso della 53, osserviamo che il test statistc cade esattamente sotto la curva del χ^2 evidenziando la validità dell'ipotesi nulla, ovvero l'assenza di discrepanze nel data sample.

Per quanto riguarda la 65, invece, notiamo che il test statistic si trova estremamente al di fuori della distribuzione, e un analogo risultato si trova per la 42, rilevando le anomalie presenti nelle ultime due run considerate.

- Ricostruzione della Rete - Slide 10

In questa slide in basso invece possiamo osservare il likelihood-fit del rapporto delle distribuzioni, in verde, confrontato con il vero rapporto delle due, mostrato in blu. Notiamo immediatamente quindi che la rete ricostruisce in modo soddisfacente il rapporto all'interno della scatola dei tempi, ovvero tra 0 e 400 ns, mentre al di fuori di questo intervallo si comporta in modo anomalo.

Lo stesso fenomeno lo possiamo notare nei tre grafici in alto, dove sono mostrati il reference, ovvero l'istogramma pieno azzurro, il data sample in blu e la ricostruzione del data sample ottenuto utilizzando il fit della rete in verde.

La causa di questo comportamento è sicuramente la semplicità della rete che abbiamo utilizzato, e ci aspettiamo che aumentando la complessità dell'architettura, e di conseguenza ad un numero maggiore di parametri liberi, la rete aumenti in potere espressivo e che possa approssimare meglio anche le regioni più critiche.

- Riassunto - Slide 11

Riassumendo brevemente, abbiamo notato che l'algoritmo riesce in modo soddifacente a comparare i data samples con il reference. Infatti, le anomalie presenti nei dati sono state rivelate correttamente ed è stato

ritornato in output un test stistic maggiore qualora la discrepanza fosse stata più evidente.

Tuttavia, non è necessariamente vero che le anomalie più evidenti siano anche quelle più critiche e pericolose da un punto di vista di malfunzionamento del detector. Per esempio, dei canali rumorosi possono causare una discrepanza considerevole, senza però prensetare un serio problema del rivelatore.

Inoltre, la forma della distribuzione dei tempi di deriva può cambiare, e quindi presentare discrepanze con il reference, anche per altri motivi al di fuori di malfunzionamenti del detector.

- Correlazioni - Slide 12

Per esempio, in questo grafico osserviamo la correlazione tra la scatola dei tempi e l'angolo di passaggio del muone attraverso il detector.

Selezionando solamente tempi di deriva originati da muoni che attraversano il detector con angoli maggiori di 20 gradi, per esempio, notiamo che il picco a sinistra si alza considerevolmente, mentre considerando muoni che arrivano quasi verticalmente, la distribuzione assume una form più uniforme.

Tali differenze nella forma della distrubuzione vengono rilevate dall'algoritmo, che ritorna un test statistc elevato evidenziando un'anomalia. Tuttavia, ciò non ha origine da un malfunzionamento del detector e quindi vorremmo che il test statistic mostrasse la validità dell'ipotesi nulla.

- Future Outlook - Slide 13

Per fare ciò, ci stiamo ponendo come primo obiettivo quello di utilizzare dei dataset multi dimensionali di osservabili correlate come input dell'algoritmo, in modo da aumentarne la flessibilità.

Inoltre, attraverso lo studio del likelihood-fit del rapporto delle distribuzioni, è possibile ricostruire quale sia il malfunzionamento del detector che ha originato tale anomalia nei dati. In questo modo si andrebbe a costruire una mappa tra le note possibili problematiche del rivelatore e un ben preciso set di traduzioni in data anomalies.

Infine, l'obiettivo sarebbe costruire un'architettura che permetta di sfruttare l'algoritmo per un monitoraggio online.

Tuttavia, l'utilizzo di reti neurali rende l'algoritmo troppo lento per essere messo online. In collaborazione con un gruppo di ricerca dell'università di Genova, quindi, stiamo testando le performance di un altro algoritmo, Falkon, che riproduce esattamente il funzionamento del nostro algoritmo, sostituendo la rete neurale con metodi kernel.

Il risultato estremamente promettente che è emerso da questi ultimi test è che il tempo di training esattamente su questi dati che ho presentato si riduce da 47 minuti a 2 secondi, aprendo le porte ad un progetto decisamente interessante.

- Fine - Slide 14

Grazie per l'attenzione.