Classificazione Interpretabile di Serie Temporali utilizzando Motifs e Discords Estratti con Matrix Profile

Relatore:
Riccardo Guidotti

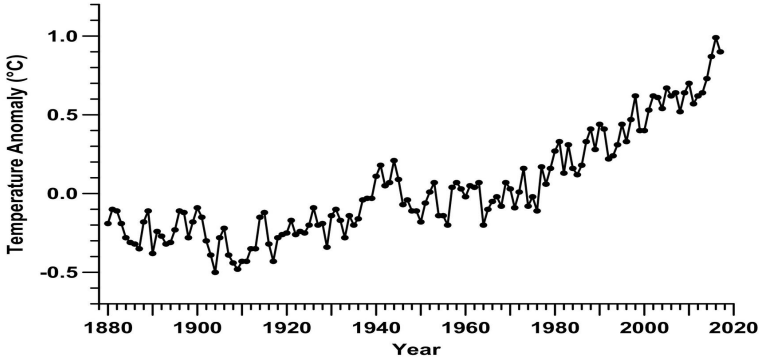
Candidato:
Matteo D'Onofrio

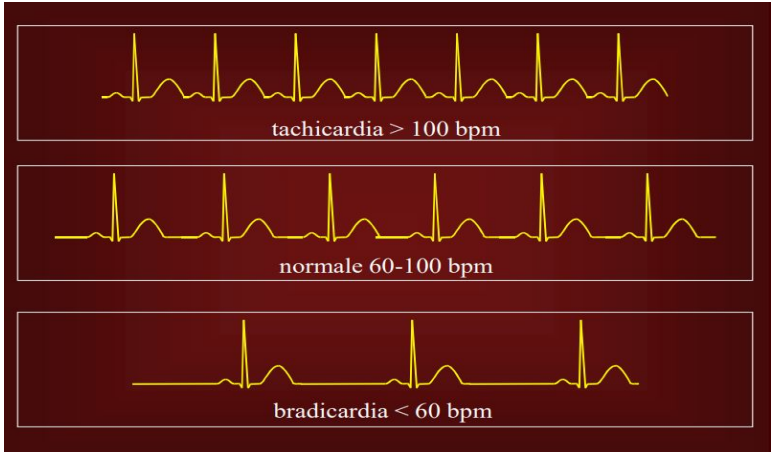
Classificazione di Serie Temporali

- Una *serie temporale* è un insieme di valori reali, ordinati rispetto al tempo, che esprimono la dinamica di un fenomeno rispetto alla variazione del tempo

- Sono presenti in economia, statistica, medicina, climatologia, epidemiologia ecc…

- Necessità di effettuare classificazioni di serie temporali, tramite metodi accurati, efficienti ed interpretabili

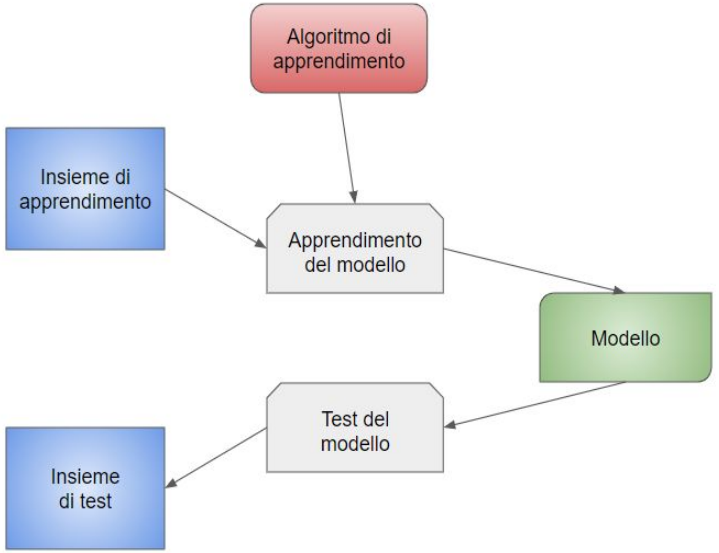




Classificazione di Serie Temporali: Alcuni Problemi

I tradizionali modelli di apprendimento supervisionato, su questo specifico tipo di classificazione mostrano empiricamente scarse prestazioni, ottenendo:

- Inefficienza data dal tempo impiegato nella fase di apprendimento
- Bassa accuratezza nella fase di test



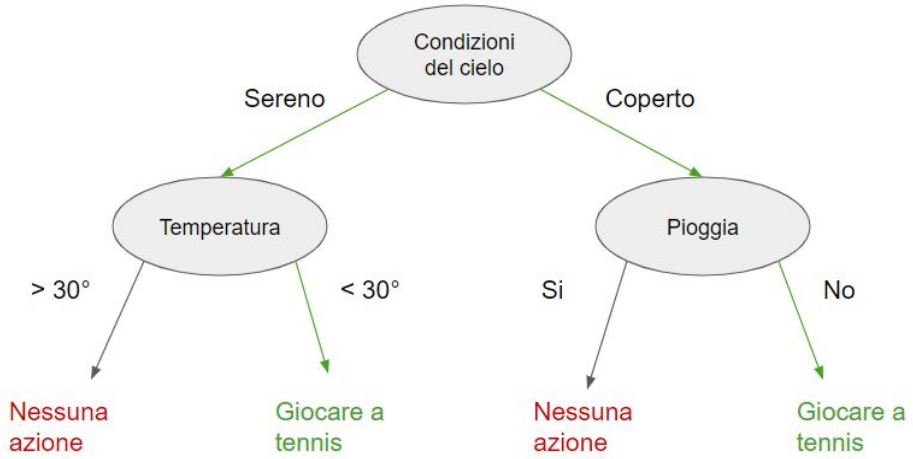
Classificazione Interpretabile

Albero di decisione (algoritmo di Hunt):

- Genera regole logiche
- Rende facilmente comprensibile il perché dei risultati ottenuti

Problemi:

- Bassa accuratezza
- Regole generate nella fase di apprendimento basate sulle informazioni fornite dai singoli valori delle serie temporali



Obiettivo della Tesi

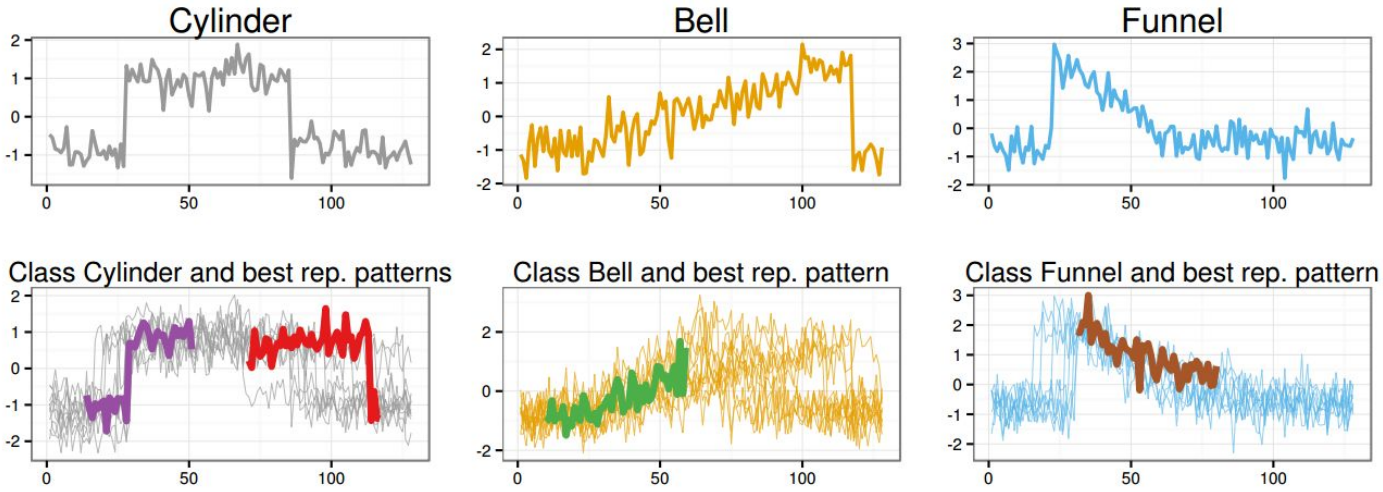
Definire ed implementare un algoritmo di apprendimento supervisionato, capace di generare un modello che classifichi le serie temporali in maniera interpretabile accurata ed efficiente

Time Series Classification based on Matrix Profile (TSCMP)

- Variante dell'albero di decisione classico (algoritmo di Hunt), garantendo interpretabilità
- Regole generate nella fase di apprendimento, basate sulle proprietà delle sottosequenze delle serie temporali, garantendo accuratezza
- Spazio di ricerca fortemente ridotto, garantendo efficienza

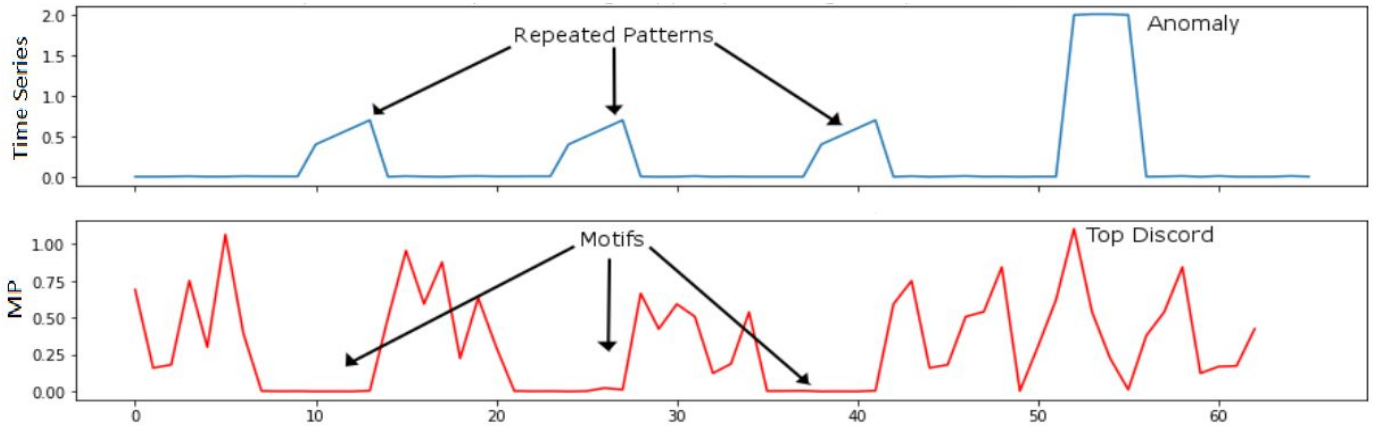
Shapelet

Una Shapelet è una sottosequenza di una serie temporale, che più di ogni altra sottosequenza, distingue ed identifica la classe di appartenenza della serie da cui è estratta



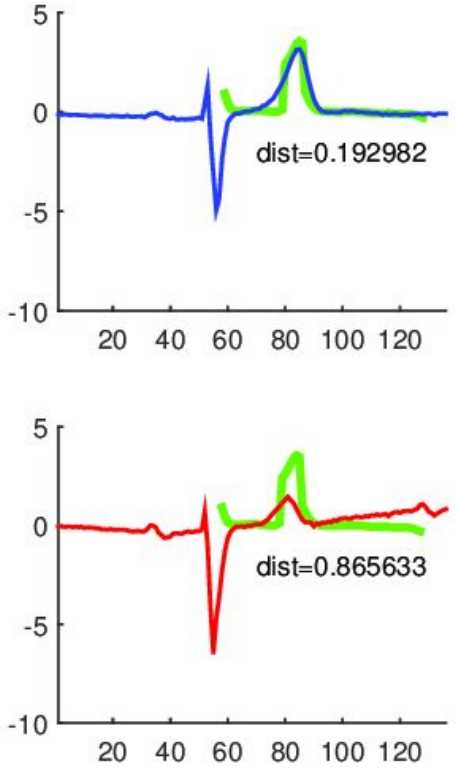
Come ottenere le Shapelet

- Fissata la dimensione delle sottosequenze, per ogni serie temporale viene calcolata la Matrix Profile (MP)
- Da ogni MP vengono estratti *k* Motifs e/o *k* Discords
- L'unione di queste sottosequenze genera l'insieme dei *candidati* Shapelet
- Il MP permette di individuare i candidati Shapelet più significativi, evitando la loro ricerca esaustiva effettuata nell'approccio classico



Apprendimento basato sulle Shapelet

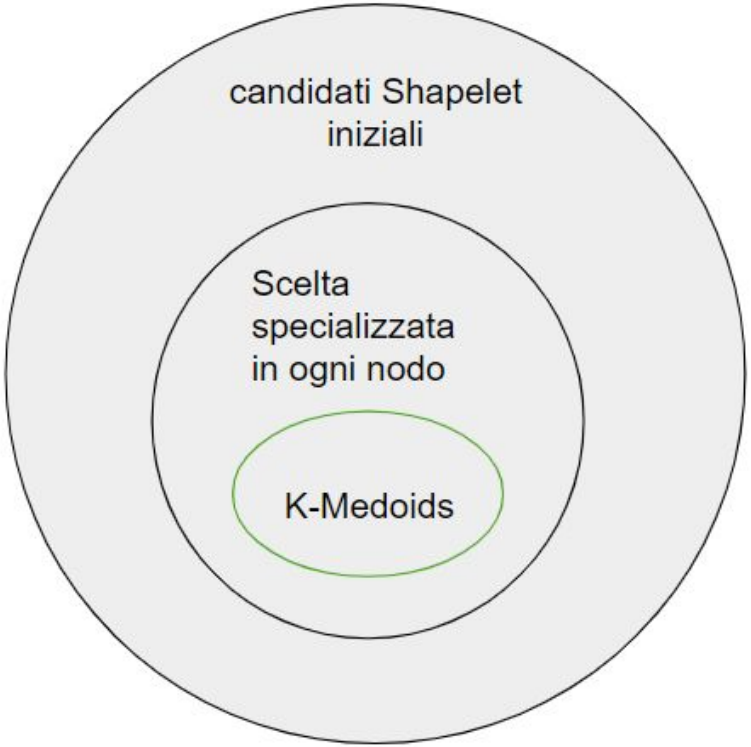
Durante la generazione delle regole, in ogni nodo dell'albero, la distanza Euclidea tra una Shapelet scelta e le serie temporali nel nodo corrente, diventa la feature che descrive la serie temporali



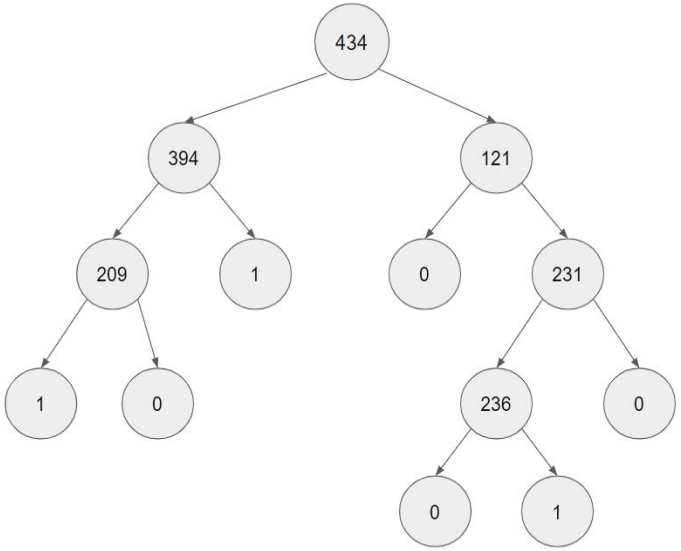
Scelta delle Shapelet durante l'apprendimento

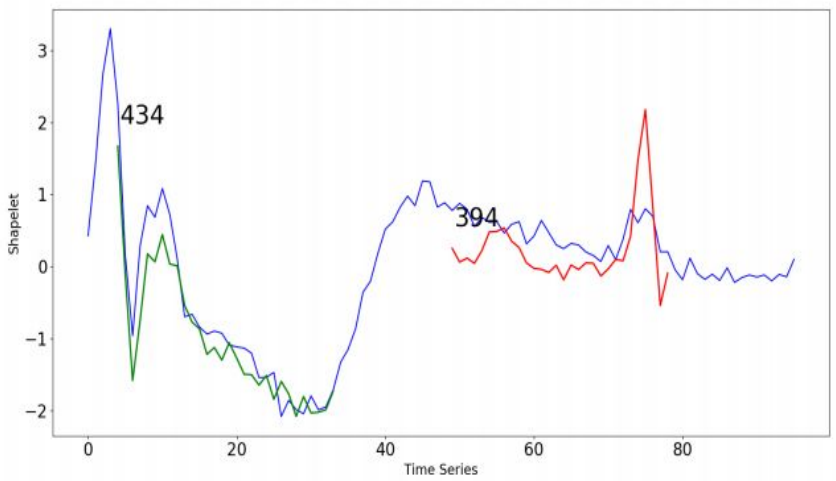
- In ogni nodo, vengono selezionati i candidati shapelet "locali", estratti dalle serie temporali presenti nel nodo correntemente utilizzato per lo split.

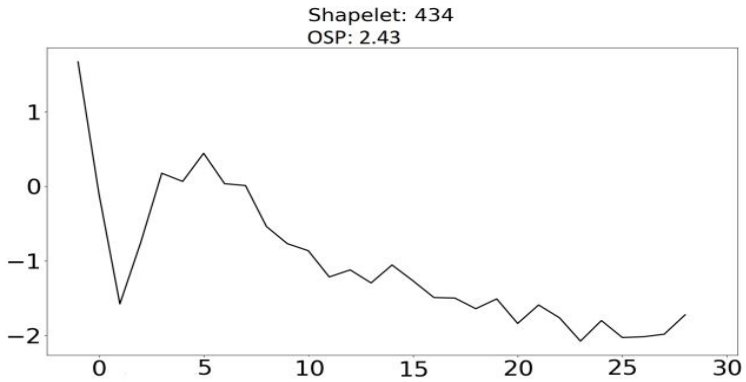
- Viene applicato l'algoritmo K-Medoids sull'insieme di Shapelet "locali" individuate selezionando così un sottoinsieme.

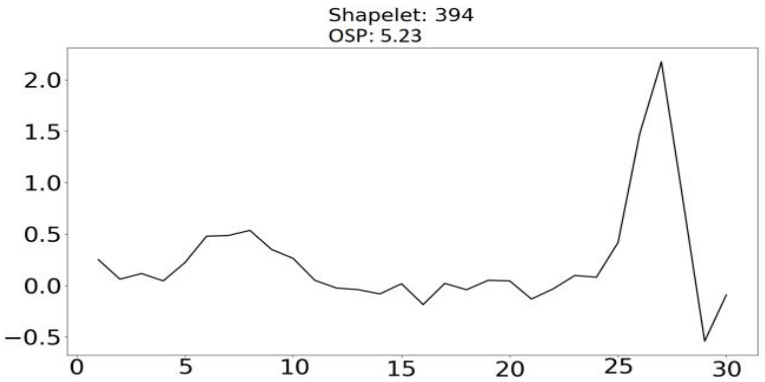


Classification Example









Esperimenti: Competitor Confrontati

ShapeletTransformation (ST)

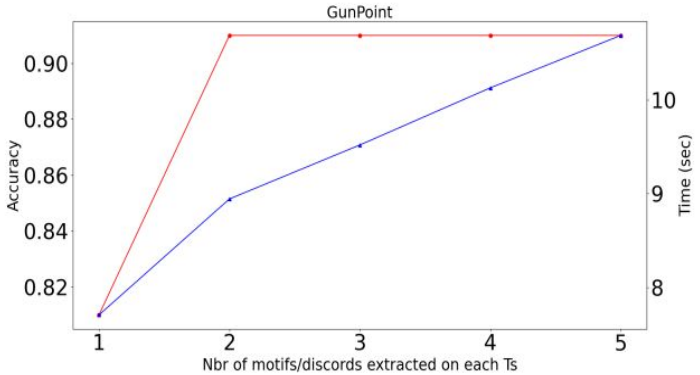
- Fissata la dimensione delle sottosequenze, per ogni serie temporale genera ogni candidato (sottosequenza) di dimensione fissata.
- Individua “globalmente” un sottoinsieme di candidati più significativi.
- Definisce un albero di classificazione in cui le regole in ogni nodo sono definite in base alla distanza tra le serie presenti nel nodo corrente e una Shapelet scelta.

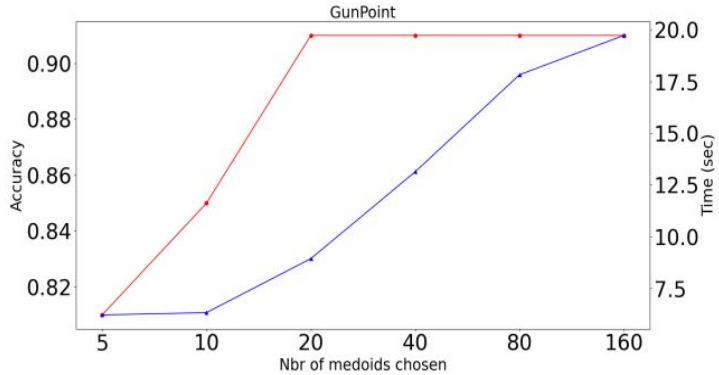
DecisionTreeClassifier (DTC)

- Riceve in input come dataset di training la distanza tra ogni serie temporale e ogni candidato individuato inizialmente dall’algoritmo TSCMP.

Esperimenti: Accuratezza ed Efficienza

Dataset	Accuracy			Tempo Esecuzione(sec)		
	TSCMP	ST	DTC	TSCMP	ST	DTC
ArrowHead	0.58	**0.64**	0.54	9.56	38.12	**8.00**
ECG200	**0.84**	0.80	0.58	18.19	25.17	**5.43**
ECG5000	**0.90**	0.89	0.36	**124.43**	1002.84	169.99
GunPoint	**0.90**	0.84	0.72	8.67	28.34	**7.23**
ItalyPowerDemand	**0.92**	0.91	0.53	9.84	**5.64**	27.06
PhalangesOutlinesCorrect	**0.70**	0.68	0.39	699.39	3777.52	**274.51**





Conclusioni

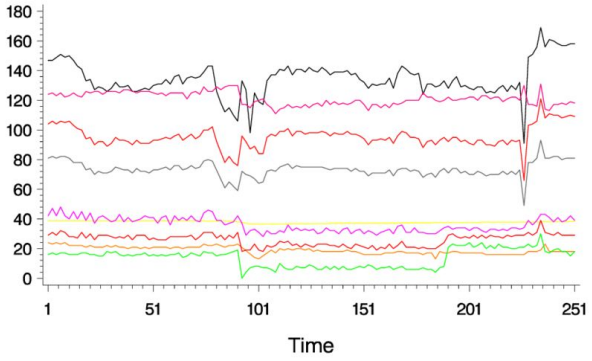
- L’uso delle Shapelet estratte delle serie temporali e la loro scelta da un insieme specializzato hanno permesso di ottenere buona accuratezza.

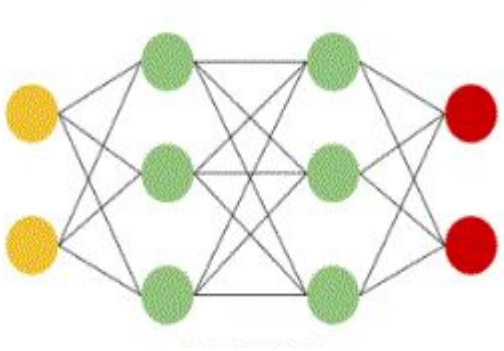
- La riduzione dello spazio di ricerca ottenuta tramite:
  - uso della Matrix Profile
  - scelta specializzata dei candidati in ogni nodo
  - applicazione del K-Medoids

  ha permesso di ottenere un’alta efficienza

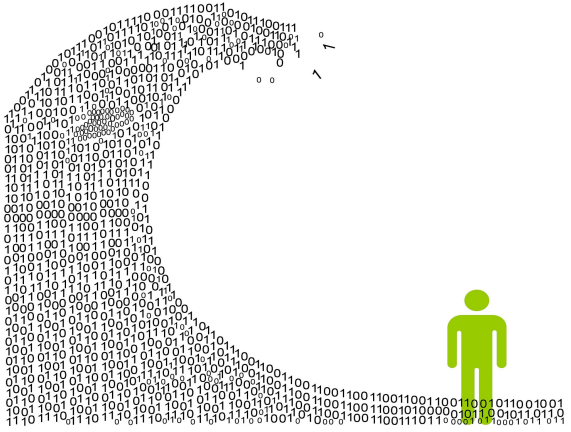
Lavori Futuri

Serie temporali multivariate





Modelli non interpretabili



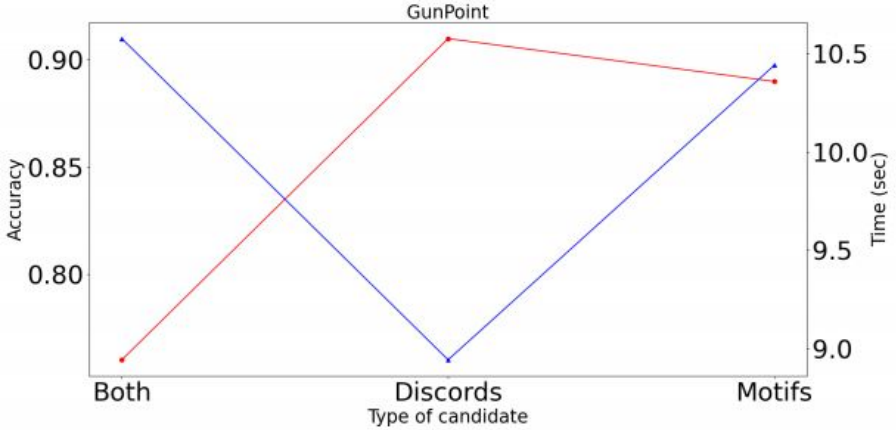
Dataset di grandi dimensioni

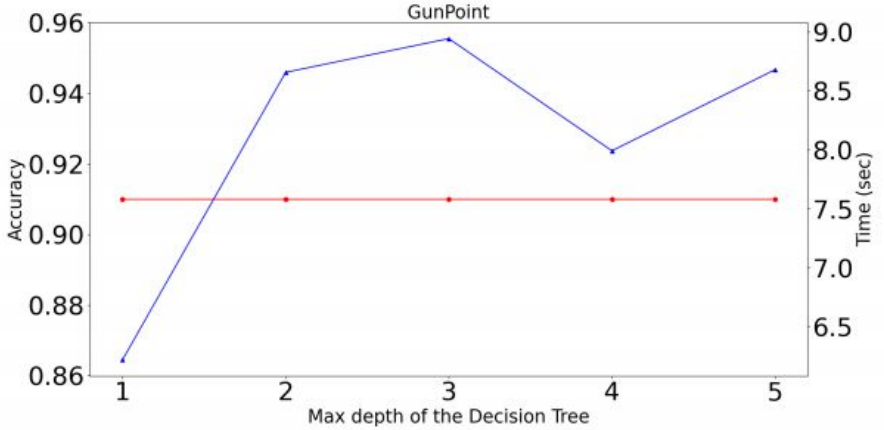
Grazie per l’attenzione

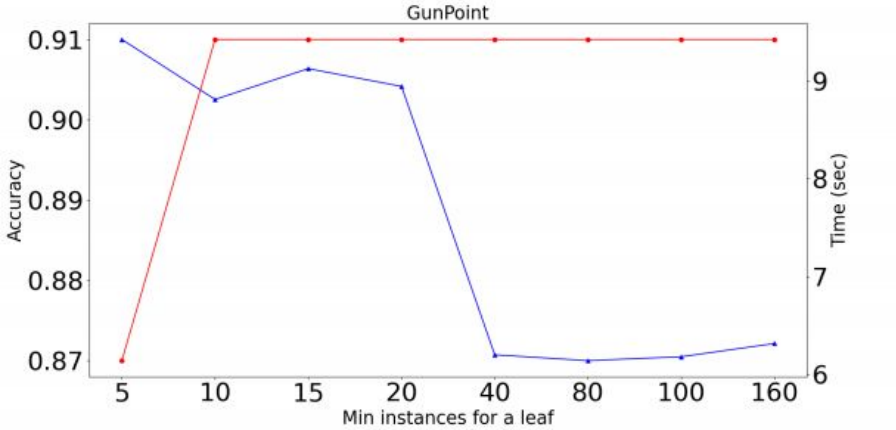
Complessità del TSCMP

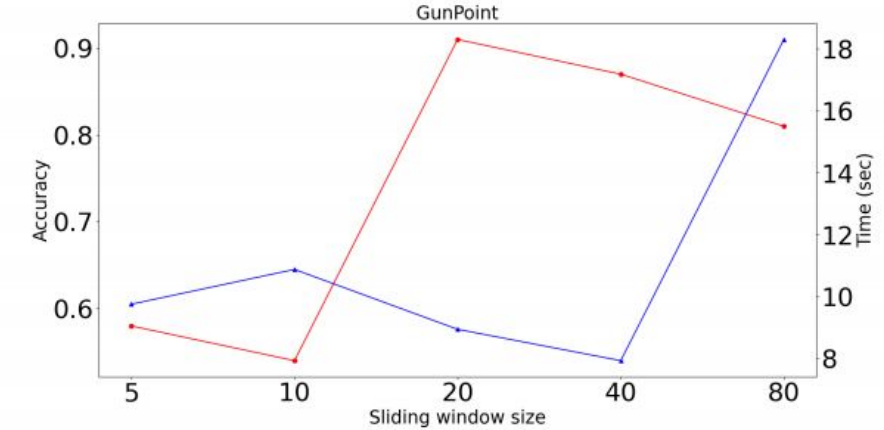
- n = #Ts, |Ts|= m, k = #M/D estratti da ogni Ts
- Ricerca candidati iniziali = O(2k (m^2) n))
- Distanze Ts-Candidati iniziali = O(n (2 k n) m log(m))
- Ricerca split = O(n nlog(n) numMedoids)
- TSCMP = O(n nlog(n) numMedoids)

Valutazione Parametri









Valutazione Parametri